Distributed Constrained Optimization by Consensus-Based Primal-Dual Perturbation Method

Tsung-Hui Chang[⋆], *Member, IEEE*, Angelia Nedić[†], *Member, IEEE*, and Anna Scaglione[‡], *Fellow, IEEE*

Revised, April 2013

Abstract

Various distributed optimization methods have been developed for solving problems which have simple local constraint sets and whose objective function is the sum of local cost functions of distributed agents in a network. Motivated by emerging applications in smart grid and distributed sparse regression, this paper studies distributed optimization methods for solving general problems which have a coupled global cost function and have inequality constraints. We consider a network scenario where each agent has no global knowledge and can access only its local mapping and constraint functions. To solve this problem in a distributed manner, we propose a consensus-based distributed primal-dual perturbation (PDP) algorithm. In the algorithm, agents employ the average consensus technique to estimate the global cost and constraint functions via exchanging messages with neighbors, and meanwhile use a local primal-dual perturbed subgradient method to approach a global optimum. The proposed PDP method not only can handle smooth inequality constraints but also non-smooth constraints such as some sparsity promoting constraints arising in sparse optimization. We prove that the proposed PDP algorithm converges to an optimal primal-dual solution of the original problem, under standard problem and network assumptions. Numerical examples illustrating the performance of the proposed algorithm for a sparse regression problem and a demand response control problem in smart grid are also presented.

Index terms– Distributed optimization, constrained optimization, average consensus, primal-dual subgradient method, regression, smart grid, demand response control

I. Introduction

Distributed optimization methods are becoming popular options for solving several engineering problems, including parameter estimation, detection and localization problems in sensor networks

The work of Tsung-Hui Chang is supported by National Science Council, Taiwan (R.O.C.), by Grant NSC 101-2218-E-011-043. The work of Angelia Nedić is supported by the NSF grants CMMI 07-42538 and CCF 11-11342. The work of Anna Scaglione is supported by NSF CCF-1011811.

[⋆]Tsung-Hui Chang is the corresponding author. Address: Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan, (R.O.C.). E-mail: tsunghui.chang@ieee.org.

[†]Angelia Nedić is with Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. E-mail: angelia@illinois.edu.

[‡]Anna Scaglione is with Department of Electrical and Computer Engineering, University of California, Davis, CA 95616, USA. E-mail: ascaglione@ucdavis.edu.

[1], [2], resource allocation problems in peer-to-peer/multi-cellular communication networks [3], [4], and distributed learning and regression problems in control [5] and machine learning [6]–[8], to name a few. In these applications, rather than pooling together all the relevant parameters that define the optimization problem, distributed agents, which have access to a local subset of such parameters, collaborate with each other to minimize a global cost function, subject to local variable constraints. Specifically, since it is not always efficient for the agents to exchange across the network the local cost and constraint functions, owing to the large size of network, time-varying network topology, energy constraints and/or privacy issues, distributed optimization methods that utilize only local information and messages exchanged between connecting neighbors have been of great interest; see [9]–[16] and references therein.

Contributions: Different from the existing works [9]–[14] where the local variable constraints are usually simple (in the sense that they can be handled via simple projection) and independent among agents, in this paper, we consider a problem formulation that has a general set of convex inequality constraints that couple all the agents' optimization variables. In addition, similar to [17], the considered problem has a global (non-separable) convex cost function that is a function of the sum of local mapping functions of the local optimization variabless. Such a problem formulation appears, for example, in the classical regression problems which have a wide range of applications. In addition, the considered formulation also arises in the demand response control and power flow control problems in the emerging smart grid systems [18]–[20]. More discussions about applications are presented in Section II-B.

In this paper, we assume that each agent knows only the local mapping function and local constraint function. To solve this problem in a distributed fashion, in this paper, we develop a novel *distributed consensus-based primal-dual perturbation (PDP)* algorithm, which combines the ideas of the primal-dual perturbed (sub-)gradient method [21], [22] and the average consensus techniques [10], [23], [24]. In each iteration of the proposed algorithm, agents exchange their local estimates of the global cost and constraint functions with their neighbors, followed by performing one-step of primal-dual variable (sub-)gradient update. Instead of using the primal-dual iterates computed at the preceding iteration as in most of the existing primal-dual subgradient based methods [15], [16], the (sub-)gradients in the proposed distributed PDP algorithm are computed based on some perturbation points which can be efficiently computed using the messages exchanged from neighbors. In particular, we provide two efficient ways to compute the perturbation points that can respectively handle the smooth and non-smooth constraint functions. More importantly, we build convergence analysis results showing that the proposed distributed PDP algorithm ensures a strong convergence of the local primal-dual iterates to a global optimal primal-dual solution of the considered problem. The proposed algorithm is applied to a distributed sparse regression problem and a distributed demand response control problem in smart grid. Numerical results for the two applications are presented to demonstrate the effectiveness of the proposed algorithm.

Related works: Distributed dual subgradient method (e.g., dual decomposition) [25] is a popular approach to solving a problem with coupled inequality constraints in a distributed manner. However, given the dual variables, this method requires the agents to globally solve the local subproblems, which may require considerable computational efforts if the local cost and constraint functions have some complex structure. Consensus-based distributed primal-dual (PD) subgradient methods have been developed recently in [15], [16] for solving a problem with an objective function which is the sum of local convex cost functions, and with global convex inequality constraints. In addition to having a different cost function from our problem formulation, the works in [15], [16] assumed that all the agents in the network have global knowledge of the inequality constraint function; the two are in sharp contrast to our formulation where a non-separable objective function is considered and each agent can access only its local constraint function. Moreover, these works adopted the conventional PD subgradient updates [26], [27] without perturbation. Numerical results will show that these methods do not perform as well as the proposed algorithm with perturbation. Another recent development is the Bregman-distance based PD subgradient method proposed in [28] for solving an epigraph formulation of a min-max problem. The method in [28], however, assumes that the Lagrangian function has a unique saddle point, in order to guarantee the convergence of the primal-dual iterates. In contrast, our proposed algorithm, which uses the perturbed subgradients, does not require such assumption.

Synopsis: Section II presents the problem formulation, applications, and a brief review of the centralized PD subgradient methods. Section III presents the proposed distributed consensus-based PDP algorithm. The assumptions and convergence analysis results are given in Section IV. Numerical results are presented in Section V. Finally, the conclusions and discussion of future extensions are drawn in Section VI.

II. Problem Formulation, Applications and Brief Review

A. Problem Formulation

We consider a network with N agents, denoted by $\mathcal{V} = \{1, \ldots, N\}$. We assume that, for all $i = 1, \ldots, N$, each agent i has a local decision variable[1] $\boldsymbol{x}_i \in \mathbb{R}^K$, a local constraint set $\mathcal{X}_i \subseteq \mathbb{R}^K$, and a local mapping function $\boldsymbol{f}_i : \mathbb{R}^K \to \mathbb{R}^M$, in which $\boldsymbol{f}_i = (f_{i1}, \ldots, f_{iM})^T$ with each $f_{im} : \mathbb{R}^K \to \mathbb{R}$ being continuous. The network cost function is given by

$$\bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \triangleq \mathcal{F}\left(\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i)\right), \tag{1}$$

where $\mathcal{F} : \mathbb{R}^M \to \mathbb{R}$ and $\bar{\mathcal{F}} : \mathbb{R}^{NK} \to \mathbb{R}$ are continuous. In addition, the agents are subject to a global inequality constraint $\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i) \preceq \boldsymbol{0}$, where $\boldsymbol{g}_i : \mathbb{R}^K \to \mathbb{R}^P$ are continuous mappings for all $i = 1, \ldots, N$; specifically, $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{iP})^T$, with each $g_{ip} : \mathbb{R}^K \to \mathbb{R}$ being continuous. The vector inequality $\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i) \preceq \boldsymbol{0}$ is understood coordinate-wise.

We assume that each agent i can access $\mathcal{F}(\cdot)$, $\boldsymbol{f}_i(\cdot)$, $\boldsymbol{g}_i(\cdot)$ and $\mathcal{X}_i$ only, for all $i = 1, \ldots, N$. Under this local knowledge constraint, the agents seek to cooperate with each other to minimize the total network cost $\bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ (or maximize the network utility $-\bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$). Mathematically, the optimization problem can be formulated as follows

$$\min_{\substack{\boldsymbol{x}_i \in \mathcal{X}_i, \\ i=1,\ldots,N}} \ \bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \ \text{ s.t. } \sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i) \preceq \boldsymbol{0}. \tag{2}$$

The goal of this paper is to develop a *distributed* algorithm for solving (2) with each agent communicating with their neighbors only.

B. Applications

In this subsection, we discuss some applications where the problem formulation (2) may arise.

Smart grid demand response and power flow control: Consider a power grid system where a retailer (e.g., the utility company) bids electricity from the power market and serves a residential/industrial neighborhood with N customers. In addition to paying for its market bid, the retailer has to pay additional cost if there is a deviation between the bid purchased in earlier market settlements and the real-time aggregate load of the customers. Any demand excess or

[1] Here, without loss of generality, we assume that all the agents have the same variable dimension K. The proposed algorithm and analysis can be easily generalized to the case with different variable dimensions.

shortfall results in a cost for the retailer that mirrors the effort to maintain the power balance. In the smart grid, thanks to the advances in communication and sensory technologies, it is envisioned that the retailer can observe the load of customers and can even control the power usage of some of the appliances (e.g., controlling the charging rate of electrical vehicles and turning ON/OFF air conditioning systems), which is known as the demand response (DR) control problem; see [29] for a recent review.

We let p_t, $t = 1, \ldots, T$, be the power bids over a time horizon of length T, and let $\psi_{i,t}(\boldsymbol{x}_i)$, $t = 1, \ldots, T$, be the load profile of customer i, where $\boldsymbol{x}_i \in \mathbb{R}^K$ contains some control variables. The structures of $\psi_{i,t}$ and $\boldsymbol{x}_i$ depend on the appliance load model. As mentioned, the retailer aims to minimize the cost caused by power imbalance, e.g., [18], [19], [29]

$$\min_{\boldsymbol{x}_1 \in \mathcal{X}_1, \ldots, \boldsymbol{x}_K \in \mathcal{X}_K} \; C_{\rm p}\left[\left(\sum_{i=1}^N \boldsymbol{\psi}_i(\boldsymbol{x}_i) - \boldsymbol{p}\right)^+\right] + C_{\rm s}\left[\left(\boldsymbol{p} - \sum_{i=1}^N \boldsymbol{\psi}_i(\boldsymbol{x}_i)\right)^+\right] \tag{3}$$

where $(x)^+ = \max\{x, 0\}$, $\mathcal{X}_i$ denotes the local control set and $C_{\rm p}, C_{\rm s} : \mathbb{R}^T \to \mathbb{R}$ denote the cost functions due to insufficient and excessive power bids, respectively. Moreover, let $\boldsymbol{p} = (P_1, \ldots, P_T)^T$ and $\boldsymbol{\psi}_i = (\psi_{i,1}, \ldots, \psi_{i,T})^T$. By defining $\boldsymbol{z} = (\sum_{i=1}^N \boldsymbol{\psi}_i(\boldsymbol{x}_i) - \boldsymbol{p})^+$ and assuming that $C_{\rm p}$ is monotonically increasing, one can write (3) as

$$\min_{\boldsymbol{x}_1 \in \mathcal{X}_1, \ldots, \boldsymbol{x}_K \in \mathcal{X}_K} \; C_{\rm p}[\boldsymbol{z}] + C_{\rm s}\left[\boldsymbol{z} - \sum_{i=1}^N \boldsymbol{\psi}_i(\boldsymbol{x}_i) + \boldsymbol{p}\right] \tag{4}$$

$$\text{s.t.} \;\; \sum_{i=1}^N \boldsymbol{\psi}_i(\boldsymbol{x}_i) - \boldsymbol{p} - \boldsymbol{z} \preceq \boldsymbol{0},$$

which belongs to the considered formulation in (2). Similar problem formulations also arise in the microgrid control problems [20], [30] where the microgrid controller requires not only to control the loads but also to control the local power generation and local power storage (i.e., power flow control), in order to maintain power balancing within the microgrid; see [30] for detailed formulations.

Distributed control methods are appealing to the smart grid application since all the agents within the systems are identical and failure of one agent would not have significant impact on the performance of the whole system [31]. Besides, it also spares the retailer/microgrid controller from the task of collecting real-time load information from the customers, which not only infringes on the customer's privacy but also is not easy for a large scale neighborhood. In Section V, the proposed distributed algorithm will be applied to a DR control problem as in (4).

Distributed regression: Regression involves modeling a response signal (e.g., observation or output of an unknown system) as a function of some regression parameters, which has wide applications, including control [5], machine learning [6], [7], data mining [32], [33] and imaging processing [7]. The goal in regression is to find the regression parameters so that the predictor function output can best represent the response signal. Let us consider a multi-agent scenario where each of the agents owns a local predictor function. Let $\boldsymbol{r} \in \mathbb{R}^M$ be a response signal that is known to all agents, and let $\boldsymbol{\phi}_i(\boldsymbol{x}_i)$ be the local predictor function at agent i, where $\boldsymbol{\phi}_i : \mathbb{R}^K \to \mathbb{R}^M$ and $\boldsymbol{x}_i$ is the regression parameter. In some applications such as distributed data mining between heterogeneous sites [32], [33] (i.e., so called vertically partitioned data [17], [34], [35]) , the agents have to generate a global data model by combining the local analysis results. In such case, the regression problem is to minimize

$$C\left(\boldsymbol{r} - \sum_{i=1}^{N} \boldsymbol{\phi}_i(\boldsymbol{x}_i)\right) \tag{5}$$

where $C : \mathbb{R}^M \to \mathbb{R}$ stands for some loss function. The regression parameters across the network may have to satisfy certain constraint. For example, it is desirable that the values in $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ are sparse, which will facilitate the storage of these local analysis results [36]. Sparsity promoting constraints, such as the one-norm constraint

$$\sum_{i=1}^{N} w_i \|\boldsymbol{x}_i\|_1 \leq k_0, \tag{6}$$

can be imposed for such purpose, were $w_i \geq 0$ are some weights and k_0 specifies the sparsity level. Note that recent works in control [37], [38] considered sparsity of some control signals and thus involve dealing with the sparsity promoting functions also. The problem formulation in (5) and (6) thus falls within the category of formulation (2). In Section V, we will also examine the proposed distributed algorithm by considering a sparse regression problem as in (5) and (6).

In addition to the above two applications, formulation (2) also encompasses the network flow control problems [39] where flow control is usually subject to capacity and flow conservation constraints; see [40] for an example which considered maximizing the network lifetime.

C. Review of Centralized PD Subgradient Method

Let us consider the following Lagrange dual problem of (2):

$$\max_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \left\{ \min_{\boldsymbol{x} \in \mathcal{X}} \ \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \right\}, \tag{7}$$

where $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T$, $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, $\boldsymbol{\lambda} \in \mathbb{R}_+^P$ ($\mathbb{R}_+^P$ is the non-negative orthant in $\mathbb{R}^P$) is the dual variable associated with the inequality constraint $\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i) \preceq \mathbf{0}$, and $\mathcal{L} : \mathbb{R}^{NK} \times \mathbb{R}_+^P \to \mathbb{R}$ is the Lagrangian function, given by

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) + \boldsymbol{\lambda}^T \left(\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i) \right). \tag{8}$$

We assume that strong duality holds for problem (2) [41]:

Assumption 1 Problem (2) is a convex problem and Slater's condition holds, i.e., there is an $(\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_N)$ that lies in the relative interior of $\mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ such that $\sum_{i=1}^N \boldsymbol{g}_i(\bar{\boldsymbol{x}}_i) \prec \mathbf{0}$.

Under such condition, one is able to handle (2) by solving its dual in (7). A classical approach along this line is the dual subgradient method [42]. Specifically, given a dual variable $\boldsymbol{\lambda}^{(k-1)}$ at iteration k, one solves the inner minimization problem

$$\boldsymbol{x}^{(k)} = \arg \min_{\boldsymbol{x} \in \mathcal{X}} \ \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}^{(k-1))}), \tag{9}$$

followed by updating the dual variable by $\boldsymbol{\lambda}^{(k)} = \big(\boldsymbol{\lambda}^{(k-1)} + a_k \ \sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)}) \big)^+$ for the outer maximization part in (7), where $a_k > 0$ is the step size. One limitation of the dual subgradient method is that the inner problem (9) needs to be globally solved, which, however, is not always easy. Even when $\bar{\mathcal{F}}(\boldsymbol{x}) = \sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i)$ for which (9) can be decomposed into N parallel subproblems, attaining the global optimum for each subproblem may still require considerable computational efforts if $\boldsymbol{f}_i(\boldsymbol{x}_i)$, $\boldsymbol{g}_i(\boldsymbol{x}_i)$ and the local set $\mathcal{X}_i$ have complex structures. One should note that the dual decomposition method [25] is exactly based on the dual subgradient method.

Another approach to dealing with (7) is the (centralized) primal-dual (PD) subgradient method [26], [43] which replaces (9) by a simple primal subgradient update. More precisely, the PD subgradient method can be described as follows. At iteration k, perform

$$\boldsymbol{x}^{(k)} = \mathcal{P}_{\mathcal{X}}(\boldsymbol{x}^{(k-1)} - a_k \ \mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)})), \tag{10a}$$

$$\boldsymbol{\lambda}^{(k)} = (\boldsymbol{\lambda}^{(k-1)} + a_k \ \mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}))^+, \tag{10b}$$

where $\mathcal{P}_{\mathcal{X}} : \mathbb{R}^{NK} \to \mathcal{X}$ is a projection function, $a_k > 0$ is the step size, and

$$\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}) \triangleq \begin{bmatrix} \mathcal{L}_{\boldsymbol{x}_1}(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}) \\ \vdots \\ \mathcal{L}_{\boldsymbol{x}_N}(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}) \end{bmatrix} = \begin{bmatrix} \nabla \boldsymbol{f}_1^T(\boldsymbol{x}_1^{(k)}) \nabla \mathcal{F}(\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})) + \nabla \boldsymbol{g}_1^T(\boldsymbol{x}_1^{(k)})\boldsymbol{\lambda}^{(k)} \\ \vdots \\ \nabla \boldsymbol{f}_N^T(\boldsymbol{x}_N^{(k)}) \nabla \mathcal{F}(\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})) + \nabla \boldsymbol{g}_N^T(\boldsymbol{x}_N^{(k)})\boldsymbol{\lambda}^{(k)} \end{bmatrix}, \tag{11a}$$

$$\mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}) \triangleq \sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)}), \tag{11b}$$

represent the subgradients of $\mathcal{L}$ at $(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})$ with respect to $\boldsymbol{x}$ and $\boldsymbol{\lambda}$, respectively. Each $\nabla \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$ is a $P \times K$ matrix with rows equal to the subgradients $\nabla g_{ip}^T(\boldsymbol{x}_i)$, $p = 1, \ldots, P$ (gradients if they are continuously differentiable), and each $\nabla \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$ is a $M \times K$ matrix with rows containing the gradients $\nabla f_{im}^T(\boldsymbol{x}_i)$, $m = 1, \ldots, M$.

The idea behind the PD subgradient method lies in the following well-known saddle-point relation, provided that the strong duality holds:

Theorem 1 (Saddle-Point Theorem) [41] *The point* $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star) \in \mathcal{X} \times \mathbb{R}_+^P$ *is a primal-dual solution pair of problems* (2) *and* (7) *if and only if there holds:*

$$\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{\lambda}) \leq \mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star) \leq \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}^\star)\ \forall \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\lambda} \succeq \boldsymbol{0}. \tag{12}$$

According to Theorem 1, if the PD subgradient method converges to a saddle point of the Lagrangian function (8), then it solves the original problem (2). Convergence properties of the PD method in (10) have been studied extensively; see, for example, [26], [27], [43]. In such methods, typically a subsequence of the sequence $(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})$ converges to a saddle point of the Lagrangian function in (8). To ensure the convergence of the whole sequence $(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})$, it is often assumed that the Lagrangian function is strictly convex in $\boldsymbol{x}$ and strictly concave in $\boldsymbol{\lambda}$, which does not hold in general however.

One of the approaches to circumventing this condition is the primal-dual perturbed (PDP) subgradient method in [21], [22]. Specifically, [21] suggests to update $\boldsymbol{x}^{(k-1)}$ and $\boldsymbol{\lambda}^{(k-1)}$ based on some perturbation points, denoted by $\hat{\boldsymbol{\alpha}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$, respectively. The PDP updates are

$$\boldsymbol{x}^{(k)} = \mathcal{P}_{\mathcal{X}}(\boldsymbol{x}^{(k-1)} - a_k\ \mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})), \tag{13a}$$

$$\boldsymbol{\lambda}^{(k)} = (\boldsymbol{\lambda}^{(k-1)} + a_k\ \mathcal{L}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}^{(k)},\boldsymbol{\lambda}^{(k-1)}))^+. \tag{13b}$$

Note that, in (13a), we have replaced $\boldsymbol{\lambda}^{(k-1)}$ by $\hat{\boldsymbol{\beta}}^{(k)}$, and, in (13b), replaced $\boldsymbol{x}^{(k-1)}$ by $\hat{\boldsymbol{\alpha}}^{(k)}$, and thus $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})$ and $\mathcal{L}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}^{(k)},\boldsymbol{\lambda}^{(k-1)})$ are perturbed subgradients. It was shown in [21] that, with carefully chosen $(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$ and the step size a_k, the primal-dual iterates in (13) converge to a saddle point of (7), without any strict convexity and concavity assumptions on $\mathcal{L}$.

There are several ways to generate the perturbation points $\hat{\boldsymbol{\alpha}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$. Our interests lie specifically on those that are computationally as efficient as the PD subgradient updates in (13). Depending on the smoothness of $\{g_{ip}\}_{p=1}^{P}$, we consider the following two methods:

Gradient Perturbation Points: A simple approach to computing the perturbation points is using the conventional gradient updates exactly as in (10), i.e.,

$$\hat{\boldsymbol{\alpha}}^{(k)} = \mathcal{P}_{\mathcal{X}}(\boldsymbol{x}^{(k-1)} - \rho_1 \ \mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)})), \tag{14a}$$

$$\hat{\boldsymbol{\beta}}^{(k)} = (\boldsymbol{\lambda}^{(k-1)} + \rho_2 \ \mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}))^{+} \tag{14b}$$

where $\rho_1 > 0$ and $\rho_2 > 0$ are constants. The PDP subgradient method thus combines (13) and (14), which involve two primal and dual subgradient updates. Even though the updates are relatively simple, this method requires smooth constraint functions g_{ip}, $p = 1, \ldots, P$.

Proximal Perturbation Points: In cases where g_{ip}, $p = 1, \ldots, P$, are non-smooth, we compute the perturbation point $\hat{\boldsymbol{\alpha}}^{(k)}$ by the following proximal gradient update[2] [44]:

$$\hat{\boldsymbol{\alpha}}^{(k)} = \arg\min_{\boldsymbol{\alpha}\in\mathcal{X}} \left\{ \sum_{i=1}^{N} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\boldsymbol{\lambda}^{(k-1)} + \frac{1}{2\rho_1}\left\| \boldsymbol{\alpha} - (\boldsymbol{x}^{(k-1)} - \rho_1 \nabla\bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}))\right\|^2 \right\} \tag{15}$$

$$= \arg\min_{\boldsymbol{\alpha}\in\mathcal{X}} \left\{ \sum_{i=1}^{N} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\boldsymbol{\lambda}^{(k-1)} + (\boldsymbol{\alpha} - \boldsymbol{x}^{(k-1)})^T \nabla\bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}) + \frac{1}{2\rho_1}\|\boldsymbol{\alpha} - \boldsymbol{x}^{(k-1)}\|^2 \right\}, \tag{16}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_N^T)^T$ and

$$\nabla\bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}) = \begin{bmatrix} \nabla \boldsymbol{f}_1^T(\boldsymbol{x}_1^{(k-1)})\nabla\mathcal{F}(\sum_{i=1}^{N}\boldsymbol{f}_1(\boldsymbol{x}_1^{(k-1)})) \\ \vdots \\ \nabla \boldsymbol{f}_N^T(\boldsymbol{x}_N^{(k-1)})\nabla\mathcal{F}(\sum_{i=1}^{N}\boldsymbol{f}_N(\boldsymbol{x}_N^{(k-1)})) \end{bmatrix}. \tag{17}$$

It is worthwhile to note that, when g_{ip}, $p = 1, \ldots, P$, are some *sparsity promoting functions* (e.g., the 1-norm, 2-norm and the nuclear norm) that often arise in sparse regression problems [7], [35], [45], the proximal perturbation point in (15) can be solved very efficiently and may even have closed-form solutions. For example, if $\boldsymbol{g}_i(\boldsymbol{\alpha}_i) = \|\boldsymbol{\alpha}_i\|_1$ for all i $(P = 1)$, and $\mathcal{X} = \mathbb{R}^{KN}$, (15) has a closed-form solution known as the soft thresholding operator [7]:

$$\hat{\boldsymbol{\alpha}}^{(k)} = (\boldsymbol{b} - \lambda^{(k-1)}\rho_1\boldsymbol{1})^{+} + (-\boldsymbol{b} - \lambda^{(k-1)}\rho_1\boldsymbol{1})^{+}, \tag{18}$$

[2] If not mentioned specifically, the norm function $\|\cdot\|$ stands for the Euclidian norm.

where $\boldsymbol{b} = \boldsymbol{x}^{(k-1)} - \rho_1 \nabla \bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)})$ and $\mathbf{1}$ is an all-one vector. To the best of our knowledge, the proximal perturbation point (15) is novel, as it has not appeared in earlier works [21], [22].

III. Proposed Consensus-Based Distributed PDP Algorithm

Our goal is to develop a distributed counterpart of the PDP subgradient method in (13). Let us recall Assumption 1 and consider the following saddle-point problem

$$\max_{\boldsymbol{\lambda} \in \mathcal{D}} \left\{ \min_{\substack{\boldsymbol{x}_i \in \mathcal{X}_i, \\ i=1,\ldots,N}} \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\lambda}) \right\} \tag{19}$$

where

$$\mathcal{D} = \left\{ \boldsymbol{\lambda} \succeq \mathbf{0} \mid \|\boldsymbol{\lambda}\| \leq D_\lambda \triangleq \frac{\bar{\mathcal{F}}(\bar{\boldsymbol{x}}) - \tilde{q}}{\gamma} + \delta \right\} \tag{20}$$

in which $\bar{\boldsymbol{x}} = (\bar{\boldsymbol{x}}_1^T, \ldots, \bar{\boldsymbol{x}}_N^T)^T$ is a Slater point of (2), $\tilde{q} = \min_{\boldsymbol{x}_i \in \mathcal{X}_i, i=1,\ldots,N} \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \tilde{\boldsymbol{\lambda}})$ is the dual function value for some arbitrary $\tilde{\boldsymbol{\lambda}} \succeq \mathbf{0}$, $\gamma = \min_{p=1,\ldots,P}\{-\sum_{i=1}^N g_{ip}(\bar{\boldsymbol{x}}_i)\}$, and $\delta > 0$ is arbitrary. It has been shown in [46] that, under Assumption 1, the optimal dual solution of (7), denoted by $\hat{\boldsymbol{\lambda}}^\star$, satisfies

$$\|\hat{\boldsymbol{\lambda}}^\star\| \leq \frac{\bar{\mathcal{F}}(\bar{\boldsymbol{x}}) - \tilde{q}}{\gamma} \tag{21}$$

and thus $\hat{\boldsymbol{\lambda}}^\star$ lies in $\mathcal{D}$. Here we consider the saddle point problem (19), instead of the original Lagrange dual problem (7), because $\mathcal{D}$ bounds the dual variable $\boldsymbol{\lambda}$ and thus also bounds the subgradient $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k)}, \boldsymbol{\lambda}^{(k)})$ in (11a). This property is important in building the convergence of the distributed algorithm to be discussed shortly. Both (7) and (19) have the same optimal dual solution $\hat{\boldsymbol{\lambda}}^\star$ and attain the same optimal objective value. One can further verify that any saddle point of (7) is also a saddle point of (19). However, to prove the converse, some conditions are needed, as given in the following proposition.

Proposition 1 (Primal-dual optimality conditions) [47] *Suppose that Assumption 1 holds. Let* $(\hat{\boldsymbol{x}}_1^\star, \ldots, \hat{\boldsymbol{x}}_N^\star, \hat{\boldsymbol{\lambda}}^\star)$ *be a saddle point of* (19). *Then* $(\hat{\boldsymbol{x}}_1^\star, \ldots, \hat{\boldsymbol{x}}_N^\star)$ *is an optimal solution for problem* (2) *if and only if*

$$\sum_{i=1}^N \boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^\star) \preceq \mathbf{0} \textit{ and } (\hat{\boldsymbol{\lambda}}^\star)^T \left(\sum_{i=1}^N \boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^\star) \right) = \mathbf{0}.$$

Proposition 1 implies that if a saddle point of (19) is primal feasible and satisfies the complementary slackness condition, then it is also a saddle point of problem (7), i.e., an optimal primal-dual solution pair of problem (2).

To have a distributed optimization algorithm for solving (19), in addition to $\boldsymbol{x}_i^{(k)}$, we let each agent i have a local copy of the dual iterate $\boldsymbol{\lambda}^{(k)}$, denoted by $\boldsymbol{\lambda}_i^{(k)}$. Moreover, each agent i owns two auxiliary variables, denoted by $\boldsymbol{y}_i^{(k)}$ and $\boldsymbol{z}_i^{(k)}$, representing respectively the local estimates of the average values of the argument function $\frac{1}{N}\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$ and of the inequality constraint function $\frac{1}{N}\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$, for all $i=1,\ldots,N$. We consider a *time-varying synchronous network* model [11], where the network of agents at time k is represented by a weighted directed graph $\mathcal{G}(k)=(\mathcal{V},\mathcal{E}(k),\boldsymbol{W}(k))$. Here $(i,j)\in\mathcal{E}(k)$ if and only if agent i can receive messages from agent j, and $\boldsymbol{W}(k)\in\mathbb{R}^{N\times N}$ is a weight matrix with each entry $[\boldsymbol{W}(k)]_{ij}$ representing a weight that agent i assigns to the incoming message on link (i,j) at time k. If $(i,j)\in\mathcal{E}(k)$, then $[\boldsymbol{W}(k)]_{ij}>0$ and $[\boldsymbol{W}(k)]_{ij}=0$ otherwise. The agents exchange messages with their neighbors (according to the network graph $\mathcal{G}(k)$) in order to achieve consensus on $\boldsymbol{\lambda}^{(k)}$, $\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$ and $\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$; while computing local perturbation points and primal-dual (sub-)gradient updates locally. Specifically, the proposed distributed consensus-based PDP method consists of the following steps at each iteration k:

1) Averaging consensus: For $i=1,\ldots,N$, each agent i sends $\boldsymbol{y}_i^{(k-1)}$, $\boldsymbol{z}_i^{(k-1)}$ and $\boldsymbol{\lambda}_i^{(k-1)}$ to all its neighbors j satisfying $(j,i)\in\mathcal{E}(k)$; it also receives $\boldsymbol{y}_j^{(k-1)}$, $\boldsymbol{z}_j^{(k-1)}$ and $\boldsymbol{\lambda}_j^{(k-1)}$ from its neighbors, and combines the received estimates, as follows:

$$\tilde{\boldsymbol{y}}_i^{(k)}=\sum_{j=1}^N[\boldsymbol{W}(k)]_{ij}\boldsymbol{y}_j^{(k-1)},\ \ \tilde{\boldsymbol{z}}_i^{(k)}=\sum_{j=1}^N[\boldsymbol{W}(k)]_{ij}\boldsymbol{z}_j^{(k-1)},\ \ \tilde{\boldsymbol{\lambda}}_i^{(k)}=\sum_{j=1}^N[\boldsymbol{W}(k)]_{ij}\boldsymbol{\lambda}_j^{(k-1)}. \tag{22}$$

2) Perturbation point computation: For $i=1,\ldots,N$, if functions g_{ip}, $p=1,\ldots,P$, are smooth, then each agent i computes the local perturbation points by

$$\boldsymbol{\alpha}_i^{(k)}=\mathcal{P}_{\mathcal{X}_i}(\boldsymbol{x}_i^{(k-1)}-\rho_1[\nabla\boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)})+\nabla\boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\tilde{\boldsymbol{\lambda}}_i^{(k)}]), \tag{23a}$$

$$\boldsymbol{\beta}_i^{(k)}=\mathcal{P}_{\mathcal{D}}(\tilde{\boldsymbol{\lambda}}_i^{(k)}+\rho_2\ N\tilde{\boldsymbol{z}}_i^{(k)}). \tag{23b}$$

Note that, comparing to (14) and (15), agent i here uses the most up-to-date estimates $N\tilde{\boldsymbol{y}}_i^{(k)}$, $N\tilde{\boldsymbol{z}}_i^{(k)}$ and $\tilde{\boldsymbol{\lambda}}_i^{(k)}$ in place of $\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k-1)})$, $\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k-1)})$ and $\boldsymbol{\lambda}^{(k-1)}$. If g_{ip}, $p=1,\ldots,P$,

Algorithm 1 Distributed Consensus-Based PDP Algorithm

```
**Given** initial variables $\boldsymbol{x}_i^{(0)} \in \mathcal{X}_i$, $\boldsymbol{\lambda}_i^{(0)} \succeq \mathbf{0}$, $\boldsymbol{y}_i^{(0)} = \boldsymbol{f}_i(\boldsymbol{x}_i^{(0)})$ and $\boldsymbol{z}_i^{(0)} = \boldsymbol{g}_i(\boldsymbol{x}_i^{(0)})$ for each agent $i$, $i = 1, \ldots, N$; Set $k = 1$.
**repeat**
    **Averaging consensus:** For $i = 1, \ldots, N$, each agent $i$ computes (22).
    **Perturbation point computation:** For $i = 1, \ldots, N$, if $\{g_{ip}\}_{p=1}^P$ are smooth, then each agent $i$ computes the local perturbation points by (23); otherwise, each agent $i$ instead computes $\boldsymbol{\alpha}_i^{(k)}$ by (24).
    **Local variable updates:** For $i = 1, \ldots, N$, each agent $i$ updates (25), (26), (27) and (28) sequentially.
    **Set** $k = k + 1$.
**until** a predefined stopping criterion (e.g., a maximum iteration number) is satisfied.
```

are non-smooth, agent i instead computes $\boldsymbol{\alpha}_i^{(k)}$ by

$$\boldsymbol{\alpha}_i^{(k)} = \arg \min_{\boldsymbol{\alpha}_i \in \mathcal{X}_i} \left\{ \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i) \tilde{\boldsymbol{\lambda}}_i^{(k)} + \frac{1}{2\rho_1} \| \boldsymbol{\alpha}_i - (\boldsymbol{x}_i^{(k-1)} - \rho_1 \nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)}) \nabla \mathcal{F}(N \tilde{\boldsymbol{y}}_i^{(k)})) \|^2 \right\}, \tag{24}$$

for $i = 1, \ldots, N$.

3) Primal-dual perturbed subgradient update: For $i = 1, \ldots, N$, each agent i updates its primal and dual variables $(\boldsymbol{x}_i^{(k)}, \boldsymbol{\lambda}_i^{(k)})$ based on the local perturbation point $(\boldsymbol{\alpha}_i^{(k)}, \boldsymbol{\beta}_i^{(k)})$:

$$\boldsymbol{x}_i^{(k)} = \mathcal{P}_{\mathcal{X}_i}(\boldsymbol{x}_i^{(k-1)} - a_k [\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)}) \nabla \mathcal{F}(N \tilde{\boldsymbol{y}}_i^{(k)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)}) \boldsymbol{\beta}_i^{(k)}]), \tag{25}$$

$$\boldsymbol{\lambda}_i^{(k)} = \mathcal{P}_{\mathcal{D}}(\tilde{\boldsymbol{\lambda}}_i^{(k)} + a_k \ \boldsymbol{g}_i(\boldsymbol{\alpha}_i^{(k)})). \tag{26}$$

4) Auxiliary variable update: For $i = 1, \ldots, N$, each agent i updates variable $\boldsymbol{y}_i^{(k)}$, $\boldsymbol{z}_i^{(k)}$ with the changes of the local argument function $\boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$ and the constraint function $\boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$:

$$\boldsymbol{y}_i^{(k)} = \tilde{\boldsymbol{y}}_i^{(k)} + \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)}) - \boldsymbol{f}_i(\boldsymbol{x}_i^{(k-1)}), \tag{27}$$

$$\boldsymbol{z}_i^{(k)} = \tilde{\boldsymbol{z}}_i^{(k)} + \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)}) - \boldsymbol{g}_i(\boldsymbol{x}_i^{(k-1)}). \tag{28}$$

Algorithm 1 summarizes the above steps. We prove that Algorithm 1 converges under proper problem and network assumptions in the next section. Readers who are interested more in numerical performance of Algorithm 1 may go directly to Section V.

IV. Convergence Analysis

Next, in Section IV-A, we present some additional assumptions on problem (2) and the network model. The main convergence results are presented in Section IV-B.

A. Assumptions

Assumption 2 (a) The sets $\mathcal{X}_i$, $i = 1, \ldots, N$, are convex and compact. In particular, for $i = 1, \ldots, N$, there is a constant $D_x > 0$ such that

$$\|\boldsymbol{x}_i\| \leq D_x \quad \forall \boldsymbol{x}_i \in \mathcal{X}_i; \tag{29}$$

(b) The functions $f_{i1}, \ldots, f_{iM}$, $i = 1, \ldots, N$, are continuously differentiable;
(c) The constraint functions $g_{i1} \ldots, g_{iP}$, $i = 1, \ldots, N$, are convex (possibly non-smooth).

Note that Assumption 2(a) and Assumption 2(b) imply that $f_{i1}, \ldots, f_{iM}$ have uniformly bounded gradients (denoted by ∇f_{im}, $m = 1, \ldots, M$) and are Lipschitz continuous, i.e., for some $L_f > 0$,

$$\max_{1\leq m\leq M} \|\nabla f_{im}(\boldsymbol{x}_i)\| \leq L_f,\ \forall \boldsymbol{x}_i \in \mathcal{X}_i \tag{30}$$

$$\max_{1\leq m\leq M} |f_{im}(\boldsymbol{x}_i) - f_{im}(\boldsymbol{y}_i)| \leq L_f \|\boldsymbol{x}_i - \boldsymbol{y}_i\| \quad \forall \boldsymbol{x}_i, \boldsymbol{y}_i \in \mathcal{X}_i. \tag{31}$$

Similarly, Assumption 2(a) and Assumption 2(c) imply that $g_{i1} \ldots, g_{iP}$ have uniformly bounded subgradients (gradients if they are continuously differentiable) and are Lipschitz continuous, i.e., for some $L_g > 0$,

$$\max_{1\leq p\leq P} \|\nabla g_{ip}(\boldsymbol{x}_i)\| \leq L_g \quad \forall \boldsymbol{x}_i \in \mathcal{X}_i, \tag{32}$$

$$\max_{1\leq p\leq P} |g_{ip}(\boldsymbol{x}_i) - g_{ip}(\boldsymbol{y}_i)| \leq L_g \|\boldsymbol{x}_i - \boldsymbol{y}_i\| \quad \forall \boldsymbol{x}_i, \boldsymbol{y}_i \in \mathcal{X}_i. \tag{33}$$

In addition, each $\boldsymbol{f}_i$ and $\boldsymbol{g}_i$ are also bounded, i.e., there exist constants $C_f > 0$ and $C_g > 0$ such that for all $i = 1, \ldots, N$,

$$\|\boldsymbol{f}_i(\boldsymbol{x}_i)\| \leq C_f, \quad \|\boldsymbol{g}_i(\boldsymbol{x}_i)\| \leq C_g, \quad \forall \boldsymbol{x}_i \in \mathcal{X}_i, \tag{34}$$

where $\|\boldsymbol{f}_i(\boldsymbol{x}_i)\| = \sqrt{\sum_{m=1}^{M} f_{im}^2(\boldsymbol{x}_i)}$ and $\|\boldsymbol{g}_i(\boldsymbol{x}_i)\| = \sqrt{\sum_{p=1}^{P} g_{ip}^2(\boldsymbol{x}_i)}$.

We also need the following assumption on the network utility costs $\mathcal{F}$ and $\bar{\mathcal{F}}$:

Assumption 3 (a) The function $\mathcal{F}$ is continuously differentiable and has bounded and Lipschitz continuous gradients, i.e., for some $G_{\mathcal{F}} > 0$ and $L_{\mathcal{F}} > 0$, we have

$$\|\nabla\mathcal{F}(\boldsymbol{x}) - \nabla\mathcal{F}(\boldsymbol{y})\| \leq G_{\mathcal{F}}\|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^M, \tag{35}$$

$$\|\nabla\mathcal{F}\left(\boldsymbol{y}\right)\| \leq L_{\mathcal{F}} \quad \forall \boldsymbol{y} \in \mathbb{R}^M; \tag{36}$$

(b) The function $\bar{\mathcal{F}}$ is convex and has Lipschitz continuous gradients, i.e., for some $G_{\bar{\mathcal{F}}} > 0$,

$$\|\nabla\bar{\mathcal{F}}(\boldsymbol{x}) - \nabla\bar{\mathcal{F}}(\boldsymbol{y})\| \leq G_{\bar{\mathcal{F}}}\|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \tag{37}$$

Note that the convexity of $\bar{\mathcal{F}}$ and Assumption 2(a) indicate that $\bar{\mathcal{F}}$ is Lipschitz continuous, i.e., for some $L_{\bar{\mathcal{F}}} > 0$,

$$\|\bar{\mathcal{F}}(\boldsymbol{x}) - \bar{\mathcal{F}}(\boldsymbol{y})\| \leq L_{\bar{\mathcal{F}}}\|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \tag{38}$$

Assumptions 2 and 3 imply that problem (2) is a convex optimization problem. In cases that g_{ip}, $p = 1, \ldots, P$, are smooth, we make the following additional assumption:

Assumption 4 The functions g_{ip}, $p = 1, \ldots, P$, are continuously differentiable and have Lipschitz continuous gradients, i.e., there exists a constant $G_g > 0$ such that

$$\max_{1\leq p\leq P}\|\nabla g_{ip}(\boldsymbol{x}_i) - \nabla g_{ip}(\boldsymbol{y}_i)\| \leq G_g\|\boldsymbol{x}_i - \boldsymbol{y}_i\| \quad \forall \boldsymbol{x}_i, \boldsymbol{y}_i \in \mathcal{X}_i. \tag{39}$$

We also have the following assumption on the network model [11], [17]:

Assumption 5 The weighted graphs $\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(k), \boldsymbol{W}(k))$ satisfy:

(a) There exists a scalar $0 < \eta < 1$ such that $[\boldsymbol{W}(k)]_{ii} > \eta$ for all i, k and $[\boldsymbol{W}(k)]_{ij} > \eta$ if $[\boldsymbol{W}(k)]_{ij} > 0$.

(b) $\boldsymbol{W}(k)$ is doubly stochastic: $\sum_{j=1}^{N}[\boldsymbol{W}(k)]_{ij} = 1$ for all i, k and $\sum_{i=1}^{N}[\boldsymbol{W}(k)]_{ij} = 1$ $\forall j, k$.

(c) There is an integer Q such that $(\mathcal{V}, \cup_{\ell=1,\ldots,Q}\mathcal{E}(k+\ell))$ is strongly connected for all k.

Assumption 5 ensures that all the agents can sufficiently and equally influence each other in a long run.

B. Main Convergence Results

Let $A_k = \sum_{\ell=1}^{k} a_\ell$, and let

$$\hat{\boldsymbol{x}}_i^{(k-1)} = \frac{1}{A_k} \sum_{\ell=1}^{k} a_\ell \; \boldsymbol{x}_i^{(\ell-1)}, \; i = 1, \ldots, N, \tag{40}$$

be the running weighted-averages of the primal iterates $\boldsymbol{x}_i^{(0)}, \ldots, \boldsymbol{x}_i^{(k-1)}$ generated by agent i until time $k-1$. Our main convergence result for Algorithm 1 is given in the following theorem:

Theorem 2 *Let Assumptions 1-5 hold, and let $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda \sqrt{P} G_g)$. Assume that the step size sequence $\{a_k\}$ is non-increasing and such that $a_k > 0$ for all $k \geq 1$, $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$. Then, for $\hat{\boldsymbol{x}}^{(k)} = (\hat{\boldsymbol{x}}_1^{(k)}, \ldots, \hat{\boldsymbol{x}}_N^{(k)})$ and $\boldsymbol{\lambda}_i^{(k)}$, $i = 1, \ldots, N$, generated by Algorithm 1 using the gradient perturbation points in* (23)*, we have*

i) The sequence $\{\hat{\boldsymbol{x}}^{(k)}\}$ converges to an optimal solution $\boldsymbol{x}^\star \in \mathcal{X}$ of problem (2)*;*

ii) The sequences $\{\boldsymbol{\lambda}_i^{(k)}\}$, $i = 1, \ldots, N$, converge to a common dual optimal solution $\boldsymbol{\lambda}^\star$ of problem (2)*.*

Theorem 2 indicates that the proposed distributed primal-dual algorithm asymptotically yields an optimal primal and dual solution pair for the original problem (2).

The same convergence result holds if the constraint functions g_{ip}, $p = 1, \ldots, P,$ are non-smooth and the perturbation points $\boldsymbol{\alpha}_i^{(k)}$ are computed following (24):

Theorem 3 *Let Assumptions 1, 2, 3, and 5 hold, and let $\rho_1 \leq 1/G_{\bar{\mathcal{F}}}$. Assume that the step size sequence $\{a_k\}$ is non-increasing and such that $a_k > 0$ for all $k \geq 1$, $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$. Let the sequences $\{\hat{\boldsymbol{x}}^{(k)}\}$ and $\{\boldsymbol{\lambda}_i^{(k)}\}$, $i = 1, \ldots, N$, be generated by Algorithm 1 using the perturbation points in* (24) *and* (23b)*. Then, $\{\hat{\boldsymbol{x}}^{(k)}\}$ and $\{\boldsymbol{\lambda}_i^{(k)}\}$, $i = 1, \ldots, N$, converge to an optimal primal solution $\boldsymbol{x}^\star \in \mathcal{X}$ and an optimal dual solution $\boldsymbol{\lambda}^\star$ of problem* (2)*, respectively.*

The proofs of Theorems 2 and 3 are presented in Appendix A and Appendix B, respectively.

Remark 1 It is worthwhile to note that when the step size a_k has the form of $a/(b+k)$ where $a > 0, b \geq 0$, one can simply consider the running average below

$$\bar{\boldsymbol{x}}^{(k)} = \frac{1}{k} \sum_{\ell=0}^{k-1} \; \boldsymbol{x}^{(\ell)} = \left(1 - \frac{1}{k}\right) \bar{\boldsymbol{x}}^{(k-1)} + \frac{1}{k} \boldsymbol{x}^{(k-1)}, \tag{41}$$

instead of the running weighted-average in (40)[3].

V. Simulation Results

In this section, we examine the efficacy of the proposed distributed PDP method (Algorithm 1) by considering the linear sparse regression problem and the demand response control problem discussed in Section II-B.

Example 1 (Linear sparse regression): Consider the linear sparse regression problem below

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\in\mathbb{R}^K} \bar{\mathcal{F}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \|\boldsymbol{r} - \sum_{i=1}^N \boldsymbol{A}_i\boldsymbol{x}_i\|^2 \tag{42a}$$

$$\text{s.t. } \sum_{i=1}^N \|\boldsymbol{x}_i\|_1 \leq k_0, \tag{42b}$$

where $\boldsymbol{r} \in \mathbb{R}^M$ and $\boldsymbol{A}_i \in \mathbb{R}^{M\times K}$, $i = 1,\ldots,N$. To generate $\boldsymbol{r}$ and $\boldsymbol{A}_i$, $i = 1,\ldots,N$, we considered a distributed image sparse decoding task [7]. Specifically, we randomly extracted 3,000 overlapping patches with dimension 8×8 from the 512×512 BARBARA image, followed by applying the K-SVD algorithm [7] to learn a dictionary $\boldsymbol{D}$ (i.e., the predictor matrix) with size 64×900 based on the extracted patches. One of the patches was added with Gaussian noise with zero mean and variance 0.5, which is then used as the response signal $\boldsymbol{r}$. Two network scenarios were considered. The first scenario contains 10 agents ($N = 10$), each of them has a regression variable $\boldsymbol{x}_i$ with dimension 10 ($K = 10$). The predictor matrices $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_{10} \in \mathbb{R}^{64\times 10}$ were obtained from the first 100 columns of $\boldsymbol{D}$. Besides, k_0 was set to 3. The second scenario has 100 agents ($N = 100$), $K = 9$, $k_0 = 10$, and $\boldsymbol{D} = [\boldsymbol{A}_1,\ldots,\boldsymbol{A}_{100}]$ where each $\boldsymbol{A}_i \in \mathbb{R}^{64\times 9}$. For both scenarios, the network graphs $\mathcal{G}$ were randomly generated. Note that, for (42), the associated proximal perturbation point in (24) has a close-form solution similar to the soft thresholding operator in (18), and thus is easy to implement. In addition to the proposed PDP method, we also implemented the recently proposed distributed (consensus-based) PD subgradient method in [15] (which does not have perturbation point) for comparison[4]. We evaluated the normalized

[3] It can be shown [47] that $\bar{\boldsymbol{x}}^{(k)}$ also satisfies (A.23) in Appendix A-B, and thus Theorems 2 and 3 also hold for $\bar{\boldsymbol{x}}^{(k)}$.

[4] While the distributed PD method in [15] is not directly applicable to (42) due to the coupled objective function, one can utilize the linear structure to show that (42) is equivalent to the following saddle point problem (by Lagrange dual)

$$\max_{\substack{\lambda\geq 0,\\ \boldsymbol{\mu}\in\mathbb{R}^M}} \left\{ \min_{\substack{\boldsymbol{x}_i\in\mathbb{R}^K\\ i=1,\ldots,N,\\ \boldsymbol{z}\in\mathbb{R}^M}} -\frac{1}{2}\|\boldsymbol{\mu}\|^2 + \boldsymbol{\mu}^T(\boldsymbol{r} - \sum_{i=1}^N \boldsymbol{A}_i\boldsymbol{x}_i) + \lambda(\sum_{i=1}^N \|\boldsymbol{x}_i\|_1 - k_0) \right\}$$

accuracy at each iteration k:

$$\text{Normalized accuracy} = \frac{\bar{\mathcal{F}}(\boldsymbol{x}^{(k)}) - \bar{\mathcal{F}}^\star}{\bar{\mathcal{F}}^\star},$$

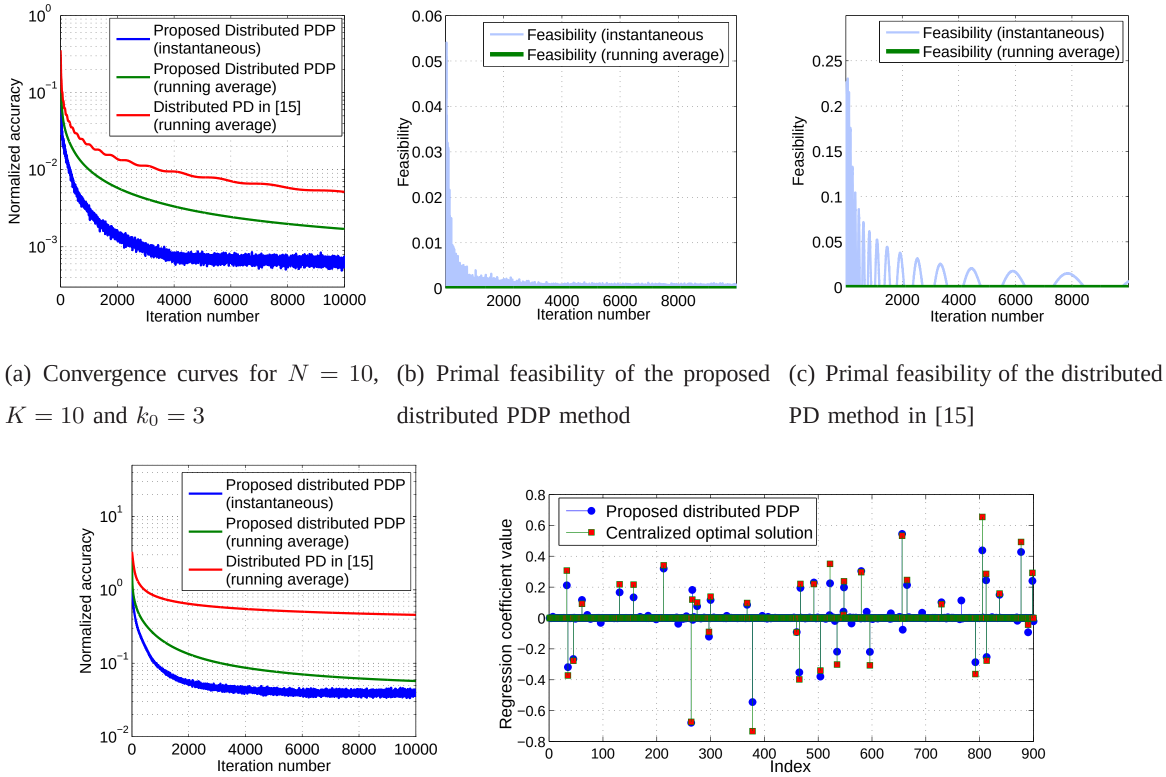
where $\bar{\mathcal{F}}^\star$ denotes the optimal value of (42) which was obtained by the (centralized) convex solver CVX [48]. In addition, we computed the primal feasibility of constraint (42b):

$$\text{Feasibility} \triangleq \frac{|(\sum_{i=1}^N \|\boldsymbol{x}_i^{(k)}\|_1 - k_0)^+|}{|\sum_{i=1}^N \|\boldsymbol{x}_i^{(k)}\|_1 - k_0|}.$$

Figure 1(a) displays the convergence curves for the first network scenario with $N = 10$, $K = 10$ and $k_0 = 3$. The step size of the distributed PD method in [15] was set to $a_k = \frac{15}{100+k}$; while the step size a_k and parameters ρ_1 and ρ_2 of the proposed distributed PDP method were set to $a_k = \frac{2}{100+k}$, $\rho_1 = 0.8$ and $\rho_2 = 1$, respectively. Note that these parameters were chosen based on cross validation so that each of the methods can respectively exhibit best convergence results. We can observe from Figure 1(a) that, for the proposed distributed PDP method, the instantaneous iterates oscillate whereas the running average iterates converge well but slower. We also see from this figure that the proposed distributed PDP method converges faster than its counterpart without perturbation in [15] (running average iterates). Figure 1(b) and Figure 1(c) respectively show the primal feasibility curves of the proposed distributed PDP method and the distributed PD method in [15]. We see that the instantaneous iterates of both methods oscillate and may not be feasible (they are nearly feasible though); whereas, for both methods, the running average iterates are always feasible.

Figure 1(d) presents the convergence curves for the second network scenario with $N = 100$, $K = 9$ and $k_0 = 10$. The step size was set to $a_k = \frac{10}{100+k}$ for the distributed PD method in [15]; while a_k, ρ_1 and ρ_2 were respectively set to $a_k = \frac{0.1}{10+k}$, $\rho_1 = 0.5$ and $\rho_2 = 0.5$ for the proposed distributed PDP method. Comparing with Figure 1(a), we first observe that the convergence speed of both methods decreases with the network size. Nevertheless, the proposed distributed PDP method still converges much faster than the method in [15]. In Figure 1(e), we further present the optimal sparse regression solution of (42) (by CVX) and that obtained by the proposed distributed PDP method (at iteration $10,000$). One can observe that the solution obtained by the proposed distributed PDP method exhibits a similar sparse pattern as the optimal solution.

to which the method in [15] can be applied.



(a) Convergence curves for $N = 10$, $K = 10$ and $k_0 = 3$

(b) Primal feasibility of the proposed distributed PDP method

(c) Primal feasibility of the distributed PD method in [15]

(d) Convergence curves for $N = 100$, $K = 9$ and $k_0 = 10$

(e) Regression coefficients

Fig. 1: Convergence curves of distributed methods for the linear sparse regression problem (42), with $N = 10$, $K = 10$ and $k_0 = 3$.

Example 2 (Smart grid demand response control): This example considers the demand response control problem presented in (3) and (4). The cost functions were set to $C_{\rm p}(\cdot) = \pi_{\rm p}\|\cdot\|^2$ and $C_{\rm s}(\cdot) = \pi_{\rm s}\|\cdot\|^2$, respectively, where $\pi_{\rm p}$ and $\pi_{\rm s}$ are some price parameters. The load profile function $\boldsymbol{\psi}_i(\boldsymbol{x}_i)$ is based on the load model in [18], which were proposed to model deferrable, non-interruptible loads such as electrical vehicle, washing machine and tumble dryer et. al. According to [18], $\boldsymbol{\psi}_i(\boldsymbol{x}_i)$ can be modeled as a linear function, i.e., $\boldsymbol{\psi}_i(\boldsymbol{x}_i) = \boldsymbol{\Psi}_i\boldsymbol{x}_i$, where $\boldsymbol{\Psi}_i \in \mathbb{R}^{T\times T}$ is a coefficient matrix composed of load profiles of appliances of customer i. The control variable $\boldsymbol{x}_i \in \mathbb{R}^T$ determines the operation scheduling of appliances of customer i. Each $\boldsymbol{x}_i$ is subject to a local constraint set $\mathcal{X}_i = \{\boldsymbol{x}_i \in \mathbb{R}^T \mid \boldsymbol{A}_i\boldsymbol{d}_i \preceq \boldsymbol{b}_i,\ \boldsymbol{l}_i \le \boldsymbol{d}_i \le \boldsymbol{u}_i\}$ due to some physical conditions and quality of service constraints [18]. The problem formulation corresponding to (3) is thus given by

$$\min_{\substack{\boldsymbol{x}_i\in\mathcal{X}_i,\\ i=1,\ldots,N}} \pi_{\rm p}\left\|\left(\sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i-\boldsymbol{p}\right)^+\right\|^2+\pi_{\rm s}\left\|\left(\boldsymbol{p}-\sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i\right)^+\right\|^2. \tag{43}$$

Analogous to (4), problem (43) can be reformulated as

$$\min_{\substack{\boldsymbol{x}_i\in\mathcal{X}_i,i=1,\ldots,N,\\ \boldsymbol{z}\succeq\boldsymbol{0}}} \quad \pi_{\rm p}\|\boldsymbol{z}\|^2+\pi_{\rm s}\left\|\boldsymbol{z}-\sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i+\boldsymbol{p}\right\|^2 \tag{44a}$$

$$\text{s.t. } \sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i-\boldsymbol{p}-\boldsymbol{z}\preceq\boldsymbol{0}, \tag{44b}$$

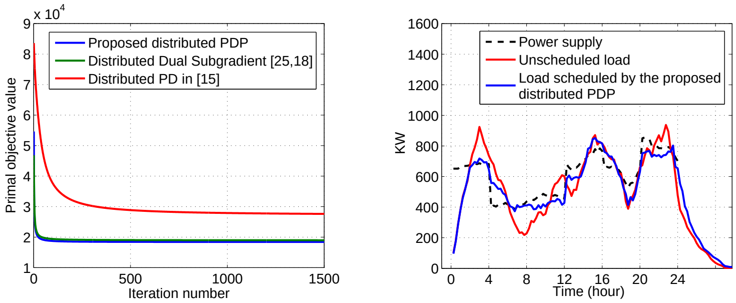
to which the proposed distributed PDP method can be applied. We consider a scenario with 400 customers ($N = 400$), and follow the same methods as in [49] to generate the power bidding $\boldsymbol{p}$ and coefficients $\boldsymbol{\Psi}_i$, $\boldsymbol{A}_i$, $\boldsymbol{b}_i$, $\boldsymbol{l}_i$, $\boldsymbol{u}_i$, $i = 1,\ldots,N$. The network graph $\mathcal{G}$ was randomly generated. The price parameters $\pi_{\rm p}$ and $\pi_{\rm s}$ were simply set to $1/N$ and $0.8/N$, respectively. In addition to the distributed PD method in [15], we also compare the proposed distributed PDP method with the distributed dual subgradient (DDS) method[5] [18], [25]. This method is based on the same idea as the dual decomposition technique [25], where, given the dual variables, each customer globally solves the corresponding inner minimization problem. The average consensus subgradient technique [10] is applied to the dual domain for distributed dual optimization.

Figure 2(a) shows the convergence curves of the three methods under test. The curves shown in this figure are the corresponding objective values in (43) of the running average iterates of the three methods. The step size of the distributed PD method in [15] was set to $a_k=\frac{15}{10+k}$ and that of the DDS method was set to $a_k=\frac{0.05}{10+k}$. For the proposed distributed PDP method, a_k, ρ_1 and ρ_2 were respectively set to $a_k=\frac{0.1}{10+k}$ and $\rho_1=\rho_2=0.001$. From this figure, we observe that the proposed distributed PDP method and the DDS method exhibit comparable convergence behavior; both methods converge within 200 iterations and outperform the distributed PD method in [15]. One should note that the DDS method is computational more expensive than the proposed

[5] One can utilize the linear structure to show that (43) is equivalent to the following saddle point problem (by Lagrange dual)

$$\max_{\substack{\boldsymbol{\lambda}\succeq\boldsymbol{0},\\ \boldsymbol{\eta}\succeq\boldsymbol{0}}}\left\{\min_{\substack{\boldsymbol{x}_i\in\mathcal{X}_i\\ i=1,\ldots,N}} -\frac{1}{4\pi_{\rm p}}\|\boldsymbol{\lambda}\|^2-\frac{1}{4\pi_{\rm s}}\|\boldsymbol{\eta}\|^2+(\boldsymbol{\lambda}-\boldsymbol{\eta})^T(\sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i-\boldsymbol{p})\right\}$$

to which the method in [15] and the DDS method [25] can be applied.



(a) Convergence curve

(b) Load profiles

Fig. 2: Numerical results for the smart grid demand response control problem (43) with 400 customers.

distributed PDP method since, in each iteration, the former requires to globally solve the inner minimization problem while the latter takes twice primal subgradient updates only.

In Figure 2(b), we further display the load profiles of the power supply, unscheduled load (without demand response control), and the load scheduled by the proposed distributed PDP method. The results were obtained by combining the proposed distributed PDP method with the certainty equivalent control (CEC) approach in [18, Algorithm 1] to handle a stochastic counterpart of problem (43). The stopping criterion was set to the maximum iteration number of 500. We can observe from this figure that the power balancing can be much improved compared to that without demand response control. Specifically, the cost in (43) is 4.49×10^4 KW for the unscheduled load whereas that of the load scheduled by the proposed distributed PDP method is 2.44×10^4 KW (45.65% reduction). The cost for the load scheduled by the distributed DDS method is slightly lower which is 2.38×10^4 KW; whereas that scheduled by the distributed PD method in [15] has a higher cost of 3.81×10^4 KW.

VI. Conclusions and Future Works

In this paper, we have presented a distributed consensus-based PDP algorithm for solving the problem formulated in (2), which has a globally coupled cost function and inequality constraints. The algorithm employs the average consensus technique and the primal-dual perturbed (sub-) gradient method. We have provided a convergence analysis showing that the proposed algorithm enables the agents across the network to achieve a global optimal primal-dual solution of the considered problem in a distributed manner. Moreover, the effectiveness of the proposed algorithm has been demonstrated by applying it to a sparse linear regression problem and a smart grid demand response control problem. In particular, the proposed algorithm is shown to have better convergence property than the distributed PD method in [15] which does not have perturbation. In addition, the proposed algorithm performs comparably with the distributed dual subgradient method [25] for the demand response control problem, even though the former is computationally cheaper.

There are several interesting research directions to pursue in the future. One direction is to extend the algorithm to asynchronous network models such as those considered in [50], [51]. The other direction is to study the convergence rate of the proposed PDP algorithm. In addition, the current practiced stopping criterion is a maximum iteration number. It would be interesting to study advanced distributed stopping criterion (e.g., based on the primal-dual optimality conditions) so that the algorithm can stop wisely in a distributed manner.

APPENDIX A
PROOF OF THEOREM 2

A. Preliminaries

Three key lemmas that will be used in the proof are presented first. The first is a deterministic version of the lemma in [52, Lemma 11, Chapter 2.2]:

Lemma 1 *Let $\{b_k\}$, $\{d_k\}$ and $\{c_k\}$ be non-negative sequences. Suppose that $\sum_{k=1}^{\infty} c_k < \infty$ and*

$$b_k \leq b_{k-1} - d_{k-1} + c_{k-1} \qquad \forall\ k \geq 1,$$

then the sequence $\{b_k\}$ converges and $\sum_{k=1}^{\infty} d_k < \infty$.

Moreover, by extending the results in [17, Theorem 4.2] and [11, Lemma 8(a)], we establish the following result on the consensus of $\{\boldsymbol{\lambda}_i^{(k)}\}$, $\{\boldsymbol{y}_i^{(k)}\}$, and $\{\boldsymbol{z}_i^{(k)}\}$ among agents.

Lemma 2 *Suppose that Assumptions 1, 2 and 5 hold. If $\{a_k\}$ is a positive, non-increasing* *sequence satisfying* $\sum_{k=1}^{\infty} a_k^2 < \infty$*, then*

$$\sum_{k=1}^{\infty} a_k \|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k)}\| < \infty, \qquad \lim_{k\to\infty} \|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k)}\| = 0,$$
$$\sum_{k=1}^{\infty} a_k \|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| < \infty, \qquad \lim_{k\to\infty} \|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| = 0,$$
$$\sum_{k=1}^{\infty} a_k \|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\| < \infty, \qquad \lim_{k\to\infty} \|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\| = 0,$$
$$\sum_{k=1}^{\infty} a_k \|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\| < \infty, \qquad \lim_{k\to\infty} \|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\| = 0,$$

for all $i = 1, \ldots, N$*, where*

$$\hat{\boldsymbol{y}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)}), \quad \hat{\boldsymbol{z}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)}), \quad \hat{\boldsymbol{\lambda}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\lambda}_i^{(k)}. \tag{A.1}$$

The proof is omitted here due to limited space. Lemma 2 implies that the local variables $\boldsymbol{\lambda}_i^{(k)}$, $\boldsymbol{y}_i^{(k)}$ and $\boldsymbol{z}_i^{(k)}$ at distributed agents will eventually achieve consensus on the values of $\hat{\boldsymbol{\lambda}}^{(k)}$ $\hat{\boldsymbol{y}}^{(k)}$ and $\hat{\boldsymbol{z}}^{(k)}$, respectively.

The next key lemma will show that the local perturbation points $\boldsymbol{\alpha}_i^{(k)}$ and $\boldsymbol{\beta}_i^{(k)}$ in (23) and (24) will also achieve some consensus asymptotically. In particular, following (14), we define

$$\hat{\boldsymbol{\alpha}}_i^{(k)} = \mathcal{P}_{\mathcal{X}_i}(\boldsymbol{x}_i^{(k-1)} - \rho_1[\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\hat{\boldsymbol{\lambda}}^{(k-1)}]), \tag{A.2a}$$
$$\hat{\boldsymbol{\beta}}^{(k)} = \mathcal{P}_{\mathcal{D}}(\hat{\boldsymbol{\lambda}}^{(k-1)} + \rho_2 \ N\hat{\boldsymbol{z}}^{(k)}), \tag{A.2b}$$

for $i = 1, \ldots, N$, as the 'centralized' counterparts of (23); similarly, following (15), we define

$$\hat{\boldsymbol{\alpha}}_i^{(k)} = \arg\min_{\boldsymbol{\alpha}_i\in\mathcal{X}_i} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\hat{\boldsymbol{\lambda}}^{(k-1)} + \frac{1}{2\rho_1}\|\boldsymbol{\alpha}_i - (\boldsymbol{x}_i^{(k-1)} - \rho_1\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)}))\|^2, \tag{A.3}$$

for $i = 1, \ldots, N$, as the centralized counterparts of the proximal perturbation point in (24). We show in Appendix C the following lemma:

Lemma 3 *Let Assumptions 2 and 3 hold. For* $\{\boldsymbol{\alpha}_i^{(k)}, \boldsymbol{\beta}_i^{(k)}\}_{i=1}^N$ *in* (23) *and* $(\hat{\boldsymbol{\alpha}}_1^{(k)}, \ldots, \hat{\boldsymbol{\alpha}}_N^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$ *in* (A.2)*, it holds that*

$$\|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{\alpha}_i^{(k)}\| \leq \rho_1 L_g\sqrt{P}\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + \rho_1 G_{\mathcal{F}} L_f\sqrt{M}N\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\|, \tag{A.4}$$
$$\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_i^{(k)}\| \leq \|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + \rho_2 N\|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\|, \tag{A.5}$$

$i = 1, \ldots, N$. *Equation* (A.4) *also holds for the proximal perturbation point* $\boldsymbol{\alpha}_i^{(k)}$ *in* (24) *and* $\hat{\boldsymbol{\alpha}}_i^{(k)}$ *in* (A.3).

Lemma 3 says that, when $\tilde{\boldsymbol{\lambda}}_i^{(k)}$, $\tilde{\boldsymbol{y}}_i^{(k)}$ and $\tilde{\boldsymbol{z}}_i^{(k)}$ at distributed agents achieve consensus, each $\boldsymbol{\alpha}_i^{(k)}$ converges to $\hat{\boldsymbol{\alpha}}_i^{(k)}$, and all the $\boldsymbol{\beta}_i^{(k)}$ converge to the common point $\hat{\boldsymbol{\beta}}^{(k)}$.

B. Proof of Theorem 2

We first show that the primal-dual iterate pairs $(\hat{\boldsymbol{x}}_1^{(k)}, \ldots, \hat{\boldsymbol{x}}_N^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ converge to a saddle point of (19), followed by showing that $(\hat{\boldsymbol{x}}_1^{(k)}, \ldots, \hat{\boldsymbol{x}}_N^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ satisfies the primal-dual optimality conditions in Proposition 1 as $k \to \infty$. The following lemma gives the basic relations for the iterates of Algorithm 1.

Lemma 4 *Let Assumptions 2 and 5 hold. Then, for any* $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T \in \mathcal{X}$ *and* $\boldsymbol{\lambda} \in \mathcal{D}$, *the following two inequalities are true:*

$$\begin{aligned}
\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|^2 \leq \|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\|^2 - 2a_k\Big(\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\beta}}^{(k)})\Big) \\
+ a_k^2 N(\sqrt{M}L_f L_{\mathcal{F}} + D_\lambda \sqrt{P} L_g)^2 + 2a_k N D_x \sqrt{M} L_f G_{\mathcal{F}} \sum_{i=1}^N \|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\| \\
+ 2a_k D_x \sqrt{P} L_g \sum_{i=1}^N \left(\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + \rho_2 N \|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}_i^{(k-1)}\|\right),
\end{aligned} \tag{A.6}$$

$$\begin{aligned}
\sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\|^2 \leq \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k-1)} - \boldsymbol{\lambda}\|^2 + 2\alpha_k\Big(\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda})\Big) + a_k^2 N C_g^2 \\
+ 2a_k(2\rho_1 D_\lambda P L_g^2 + C_g)\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + 4\rho_1 N D_\lambda G_{\mathcal{F}} \sqrt{PM} L_g L_f a_k \|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}_i^{(k-1)}\|.
\end{aligned} \tag{A.7}$$

Proof of Lemma 4: By (25), the non-expansiveness of projection [42], the subgradient boundedness in (30), (32) and (36), and by (20), one can show that, for any $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathcal{X}$,

$$\begin{aligned}
\sum_{i=1}^N \|\boldsymbol{x}_i^{(k)} - \boldsymbol{x}_i\|^2 &= \sum_{i=1}^N \|\mathcal{P}_{\mathcal{X}_i}\left(\boldsymbol{x}_i^{(k-1)} - a_k\left[\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)}) \nabla \mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\boldsymbol{\beta}_i^{(k)}\right]\right) - \boldsymbol{x}_i\|^2 \\
&\leq \sum_{i=1}^N \|\boldsymbol{x}_i^{(k-1)} - \boldsymbol{x}_i\|^2 + a_k^2 N(\sqrt{M}L_f L_{\mathcal{F}} + D_\lambda \sqrt{P} L_g)^2 \\
&\quad -2a_k \sum_{i=1}^N (\boldsymbol{x}_i^{(k-1)} - \boldsymbol{x}_i)^T (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)}) \nabla \mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\boldsymbol{\beta}_i^{(k)}).
\end{aligned} \tag{A.8}$$

The last term in (A.8) can be further bounded as follows:

$$
\begin{aligned}
&-2a_k\sum_{i=1}^{N}(\boldsymbol{x}_i^{(k-1)}-\boldsymbol{x}_i)^T(\nabla\boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)})+\nabla\boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\boldsymbol{\beta}_i^{(k)})\\
&=-2a_k\sum_{i=1}^{N}(\boldsymbol{x}_i^{(k-1)}-\boldsymbol{x}_i)^T(\nabla\boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}_i^{(k-1)})+\nabla\boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\hat{\boldsymbol{\beta}}^{(k)})\\
&\qquad\qquad-2a_k\sum_{i=1}^{N}(\boldsymbol{x}_i^{(k-1)}-\boldsymbol{x}_i)^T\nabla\boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})(\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)})-\nabla\mathcal{F}(N\hat{\boldsymbol{y}}_i^{(k-1)}))\\
&\qquad\qquad-2a_k\sum_{i=1}^{N}(\boldsymbol{x}_i^{(k-1)}-\boldsymbol{x}_i)^T\nabla\boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})(\boldsymbol{\beta}_i^{(k)}-\hat{\boldsymbol{\beta}}^{(k)})\\
&\leq-2a_k\sum_{i=1}^{N}(\boldsymbol{x}_i^{(k-1)}-\boldsymbol{x}_i)^T(\nabla\boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}_i^{(k-1)})+\nabla\boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\hat{\boldsymbol{\beta}}^{(k)})\\
&\qquad+2a_kND_x\sqrt{M}L_fG_{\mathcal{F}}\sum_{i=1}^{N}\|\tilde{\boldsymbol{y}}_i^{(k)}-\hat{\boldsymbol{y}}^{(k-1)}\|+2a_kD_x\sqrt{P}L_g\sum_{i=1}^{N}\|\boldsymbol{\beta}_i^{(k)}-\hat{\boldsymbol{\beta}}^{(k)}\|\\
&\leq-2a_k(\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})-\mathcal{L}(\boldsymbol{x},\hat{\boldsymbol{\beta}}^{(k)}))\\
&\quad+2a_kND_x\sqrt{M}L_fG_{\mathcal{F}}\sum_{i=1}^{N}\|\tilde{\boldsymbol{y}}_i^{(k)}-\hat{\boldsymbol{y}}^{(k-1)}\|+2a_kD_x\sqrt{P}L_g\sum_{i=1}^{N}\|\boldsymbol{\beta}_i^{(k)}-\hat{\boldsymbol{\beta}}^{(k)}\|, \qquad \text{(A.9)}
\end{aligned}
$$

where the first inequality is obtained by the subgradient boundedness, the Lipschitz continuity of $\nabla\mathcal{F}$, and the compactness of $\mathcal{X}_i$ and $\mathcal{D}$ (cf. (11a), (29), (35), (30), (32)), and the second inequality is due to the definition of the subgradient of a convex function, i.e., $\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})+(\boldsymbol{x}-\boldsymbol{x}^{(k-1)})^T\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})\leq\mathcal{L}(\boldsymbol{x},\hat{\boldsymbol{\beta}}^{(k)})\ \forall\boldsymbol{x}\in\mathcal{X}$. By combining (A.8) and (A.9) and applying (A.5) in Lemma 3, we obtain (A.6).

By using (26) and a line of analysis similar to that of the proof of (A.6), we can obtain, for any $\boldsymbol{\lambda}\in\mathcal{D}$,

$$
\begin{aligned}
\sum_{i=1}^{N}\|\boldsymbol{\lambda}_i^{(k)}-\boldsymbol{\lambda}\|^2&\leq\sum_{i=1}^{N}\|\tilde{\boldsymbol{\lambda}}_i^{(k)}-\boldsymbol{\lambda}\|^2+a_k^2\sum_{i=1}^{N}\|\boldsymbol{g}_i(\boldsymbol{\alpha}_i^{(k)})\|^2+2a_k\sum_{i=1}^{N}(\tilde{\boldsymbol{\lambda}}_i^{(k)}-\boldsymbol{\lambda})^T\boldsymbol{g}_i(\boldsymbol{\alpha}_i^{(k)})\\
&\leq\sum_{j=1}^{N}\|\boldsymbol{\lambda}_j^{(k-1)}-\boldsymbol{\lambda}\|^2+a_k^2NC_g^2+2a_k\sum_{i=1}^{N}(\tilde{\boldsymbol{\lambda}}_i^{(k)}-\boldsymbol{\lambda})^T\boldsymbol{g}_i(\boldsymbol{\alpha}_i^{(k)}), \qquad \text{(A.10)}
\end{aligned}
$$

where the last inequality is due to the boundedness of the function values (cf. (34)). We can bound the last term as

$$
\begin{aligned}
&2a_k\sum_{i=1}^{N}(\hat{\boldsymbol{\lambda}}^{(k-1)}-\boldsymbol{\lambda}+\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)})^T\boldsymbol{g}_i(\boldsymbol{\alpha}_i^{(k)})\\
&=2a_k(\hat{\boldsymbol{\lambda}}^{(k-1)}-\boldsymbol{\lambda})^T\left(\sum_{i=1}^{N}\boldsymbol{g}_i(\hat{\boldsymbol{\alpha}}_i^{(k)})\right)\\
&\qquad+2a_k\sum_{i=1}^{N}(\hat{\boldsymbol{\lambda}}^{(k-1)}-\boldsymbol{\lambda})^T(\boldsymbol{g}_i(\boldsymbol{\alpha}_i^{(k)})-\boldsymbol{g}_i(\hat{\boldsymbol{\alpha}}_i^{(k)}))+2a_k\sum_{i=1}^{N}(\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)})^T\boldsymbol{g}_i(\boldsymbol{\alpha}_i^{(k)})\\
&\le 2a_k(\hat{\boldsymbol{\lambda}}^{(k-1)}-\boldsymbol{\lambda})^T\left(\sum_{i=1}^{N}\boldsymbol{g}_i(\hat{\boldsymbol{\alpha}}_i^{(k)})\right)\\
&\qquad+4D_\lambda\sqrt{P}L_g a_k\sum_{i=1}^{N}\|\boldsymbol{\alpha}_i^{(k)}-\hat{\boldsymbol{\alpha}}_i^{(k)})\|+2C_g a_k\sum_{i=1}^{N}\|\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)}\|. \qquad \text{(A.11)}
\end{aligned}
$$

where in the last inequality, we have used $\tilde{\boldsymbol{\lambda}}_i^{(k)},\boldsymbol{\lambda}\in\mathcal{D}$ and the compactness of $\mathcal{D}$ (see (20)), and the Lipschitz continuity of $\boldsymbol{g}_i$ (cf. (33)). Note that, since $\mathcal{L}$ is linear in λ, we have

$$
\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\boldsymbol{\lambda})=\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)})+(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}^{(k-1)})^T\mathcal{L}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)}). \qquad \text{(A.12)}
$$

where $\hat{\boldsymbol{\alpha}}^{(k)}=((\hat{\boldsymbol{\alpha}}_1^{(k)})^T,\ldots,(\hat{\boldsymbol{\alpha}}_N^{(k)})^T)^T$ and $\mathcal{L}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)})=\sum_{i=1}^{N}\boldsymbol{g}_i(\hat{\boldsymbol{\alpha}}_i^{(k)})$. By combining (A.10), (A.11), (A.12) and (A.4) in Lemma 3, we obtain (A.7). ■

We also need the following lemma which characterizes the relation between the primal-dual iterates $(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\lambda}}^{(k-1)})$ and the centralized perturbation points $(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\beta}}^{(k)})$ in (A.2):

Lemma 5 *Let Assumptions 2, 3 and 4 hold. For the gradient perturbation points* $(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\beta}}^{(k)})$ *in* (A.2), *it holds true that*

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})-\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)})\\
&\qquad\ge\left(\frac{1}{\rho_1}-(G_{\bar{\mathcal{F}}}+D_\lambda\sqrt{P}G_g)\right)\|\boldsymbol{x}^{(k-1)}-\hat{\boldsymbol{\alpha}}^{(k)}\|^2+\frac{1}{\rho_2}\|\hat{\boldsymbol{\lambda}}^{(k-1)}-\hat{\boldsymbol{\beta}}^{(k)}\|^2. \qquad \text{(A.13)}
\end{aligned}
$$

Moreover, let $\rho_1\le 1/(G_{\bar{\mathcal{F}}}+D_\lambda\sqrt{P}G_g)$, *and suppose that* $\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})-\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)})\to 0$ *and* $(\boldsymbol{x}^{(k-1)},\boldsymbol{\lambda}^{(k-1)})$ *converges to some limit point* $(\hat{\boldsymbol{x}}^\star,\hat{\boldsymbol{\lambda}}^\star)\in\mathcal{X}\times\mathcal{D}$ *as* $k\to\infty$. *Then* $(\hat{\boldsymbol{x}}^\star,\hat{\boldsymbol{\lambda}}^\star)$ *is a saddle point of* (19).

The proof is presented in Appendix D. By Lemmas 2, 4 and 5, $(\boldsymbol{x}^{(k)},\hat{\boldsymbol{\lambda}}^{(k)})$ converges to a saddle point of (19):

Lemma 6 *Let Assumptions 2-5 hold, and let* $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda \sqrt{P} G_g)$. *Assume that the step size* $a_k > 0$ *is a non-increasing sequence satisfying* $\sum_{k=1}^{\infty} a_k = \infty$ *and* $\sum_{k=1}^{\infty} a_k^2 < \infty$. *Then*

$$\lim_{k\to\infty} \|\boldsymbol{x}_i^{(k)} - \hat{\boldsymbol{x}}_i^\star\| = 0, \quad i = 1, \ldots, N, \qquad \lim_{k\to\infty} \|\hat{\boldsymbol{\lambda}}^{(k)} - \hat{\boldsymbol{\lambda}}^\star\| = 0, \tag{A.14}$$

$$\lim_{k\to\infty} \|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\| = 0, \qquad \lim_{k\to\infty} \|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\| = 0, \tag{A.15}$$

where $\hat{\boldsymbol{x}}^\star = ((\hat{\boldsymbol{x}}_1^\star)^T, \ldots, (\hat{\boldsymbol{x}}_N^\star)^T)^T \in \mathcal{X}$ *and* $\hat{\boldsymbol{\lambda}}^\star \in \mathcal{D}$ *form a saddle point of problem* (19).

Proof of Lemma 6: By the compactness of the set $\mathcal{X}$ and the continuity of the functions $\bar{\mathcal{F}}$ and $\boldsymbol{g}_i$, problem (2) has a solution. By Assumption 1, the dual problem also has a solution. By construction of the set $\mathcal{D}$ in (20), all dual optimal solutions are contained in the set $\mathcal{D}$. We let $\boldsymbol{x}^\star = ((\boldsymbol{x}_1^\star)^T, \ldots, (\boldsymbol{x}_N^\star)^T)^T \in \mathcal{X}$ and $\boldsymbol{\lambda}^\star \in \mathcal{D}$ be an arbitrary saddle point of (19), and we apply Lemma 4 with $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T = \boldsymbol{x}^\star$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\star$. By summing (A.6) and (A.7), we obtain the following inequality

$$\begin{aligned}(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\|^2) &\leq (\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k-1)} - \boldsymbol{\lambda}^\star\|^2) \\ &+ \tilde{c}_k - 2a_k\Big(\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\beta}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) + \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^\star)\Big),\end{aligned} \tag{A.16}$$

where

$$\begin{aligned}\tilde{c}_k &\triangleq a_k^2 N[(\sqrt{M} L_f L_{\mathcal{F}} + D_\lambda \sqrt{P} L_g)^2 + C_g^2] \\ &+ 2[D_x \sqrt{P} L_g + C_g + 2\rho_1 P D_\lambda L_g^2] \sum_{i=1}^N (a_k \|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\|) + 2N\sqrt{M} L_f G_{\mathcal{F}}(D_x \\ &+ 2\rho_1 D_\lambda \sqrt{P} L_g) \sum_{i=1}^N (a_k \|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\|) + 2N\rho_2 D_x \sqrt{P} L_g \sum_{i=1}^N (a_k \|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\|).\end{aligned} \tag{A.17}$$

First of all, by Theorem 1, we have

$$\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) \geq 0, \qquad \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \hat{\boldsymbol{\beta}}^{(k)}) \geq 0,$$

implying that $\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\beta}^{(k)}) \geq 0$. Hence we deduce from (A.17) that

$$\begin{aligned}(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\|^2) &\leq (\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k-1)} - \boldsymbol{\lambda}^\star\|^2) \\ &+ \tilde{c}_k - 2a_k(\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})).\end{aligned} \tag{A.18}$$

Secondly, by $\sum_{k=1}^{\infty} a_k^2 < \infty$ and by Lemma 2, we see that all the four terms in $\tilde{c}_k$ are summable over k, and thus $\sum_{k=1}^{\infty} \tilde{c}_k < \infty$. Thirdly, by Lemma 5 and under the premise of $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda \sqrt{P} G_g)$, we have $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \geq 0$. Therefore, by applying Lemma 1 to relation (A.18), we conclude that the sequence $\{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^{N} \|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\|^2\}$ converges for any saddle point $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, and it holds that $\sum_{k=1}^{\infty} a_k \left(\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})\right) < \infty$. Because $\sum_{k=1}^{\infty} a_k = \infty$, the preceding relation implies that

$$\liminf_{k \to \infty} \mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) = 0. \tag{A.19}$$

Equation (A.19) implies that there exists a subsequence $\ell_1, \ell_2, \ldots$ such that

$$\mathcal{L}(\boldsymbol{x}^{(\ell_k-1)}, \hat{\boldsymbol{\beta}}^{(\ell_k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(\ell_k)}, \hat{\boldsymbol{\lambda}}^{(\ell_k-1)}) \to 0 \text{ as } k \to \infty. \tag{A.20}$$

According to Lemma 5, the above equation indicates that

$$\lim_{k \to \infty} \|\boldsymbol{x}^{(\ell_k-1)} - \hat{\boldsymbol{\alpha}}^{(\ell_k)}\| = 0, \qquad \lim_{k \to \infty} \|\hat{\boldsymbol{\lambda}}^{(\ell_k-1)} - \hat{\boldsymbol{\beta}}^{(\ell_k)}\| = 0. \tag{A.21}$$

Moreover, because $\{(\boldsymbol{x}^{(\ell_k-1)}, \hat{\boldsymbol{\lambda}}^{(\ell_k-1)})\} \subset \mathcal{X} \times \mathcal{D}$ is a bounded sequence, there must exist a limit point, say $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \in \mathcal{X} \times \mathcal{D}$, such that

$$\boldsymbol{x}^{(\ell_k-1)} \to \hat{\boldsymbol{x}}^\star, \qquad \hat{\boldsymbol{\lambda}}^{(\ell_k-1)} \to \hat{\boldsymbol{\lambda}}^\star, \text{ as } k \to \infty. \tag{A.22}$$

Under the premise of $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda \sqrt{P} G_g)$, and by (A.20) and (A.22), we obtain from Lemma 5 that $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \in \mathcal{X} \times \mathcal{D}$ is a saddle point of (19). Moreover, because

$$\|\boldsymbol{x}^{(\ell_k)} - \hat{\boldsymbol{x}}^\star\|^2 + \sum_{i=1}^{N} \|\boldsymbol{\lambda}_i^{(\ell_k)} - \hat{\boldsymbol{\lambda}}^\star\|^2 \leq \|\boldsymbol{x}^{(\ell_k)} - \hat{\boldsymbol{x}}^\star\|^2 + \sum_{i=1}^{N} (\|\boldsymbol{\lambda}_i^{(\ell_k)} - \hat{\boldsymbol{\lambda}}^{(\ell_k)}\| + \|\hat{\boldsymbol{\lambda}}^{(\ell_k)} - \hat{\boldsymbol{\lambda}}^\star\|)^2,$$

we obtain from Lemma 2 and (A.22) that the sequence $\{\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}^\star\|^2 + \sum_{i=1}^{N} \|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^\star\|^2\}$ has a limit value equal to zero. Since the sequence $\{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^{N} \|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\|^2\}$ converges for any saddle point of (19), we conclude that $\{\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}^\star\|^2 + \sum_{i=1}^{N} \|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^\star\|^2\}$ in fact converges to zero, and therefore (A.14) is proved. Finally, relation (A.15) can also be obtained by (A.14), (A.18) and (A.13), provided that $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda \sqrt{P} G_g)$. ■

According to [47, Lemma 3], if $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^\star$ as $k \to \infty$, then its weighted running average $\boldsymbol{x}^{(k)}$ defined in (40) also converges to $\boldsymbol{x}^\star$ as $k \to \infty$. The next lemma further shows that $\hat{\boldsymbol{x}}^{(k)}$ together with $\hat{\boldsymbol{\lambda}}^{(k)}$ asymptotically satisfy the optimality conditions given by Proposition 1.

Lemma 7 *Under the assumptions of Lemma 6, it holds*

$$\lim_{k\to\infty}\left\|\left(\sum_{i=1}^{N}\boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^{(k)})\right)^{+}\right\| = 0, \qquad \lim_{k\to\infty}(\hat{\boldsymbol{\lambda}}^{(k)})^T\left(\sum_{i=1}^{N}\boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^{(k)})\right) = 0. \tag{A.23}$$

Proof of Lemma 7: By (A.7) in Lemma 4 and the fact of $\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\boldsymbol{\lambda}) = \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)}) + (\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^{(k-1)})^T\mathcal{L}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)})$, we have

$$(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^{(k-1)})^T\boldsymbol{g}(\hat{\boldsymbol{\alpha}}^{(k)}) \leq \frac{\bar{c}_k}{2a_k} + \frac{1}{2a_k}\left(\sum_{j=1}^{N}\|\boldsymbol{\lambda}_j^{(k-1)} - \boldsymbol{\lambda}\|^2 - \sum_{i=1}^{N}\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\|^2\right), \tag{A.24}$$

where $\boldsymbol{g}(\hat{\boldsymbol{\alpha}}^{(k)}) = \sum_{i=1}^{N}\boldsymbol{g}_i(\hat{\boldsymbol{\alpha}}_i^{(k)})$ and

$$\bar{c}_k \triangleq a_k^2 N C_g^2 + 2a_k(2\rho_1 D_\lambda P L_g^2 + C_g)\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + 4\rho_1 N D_\lambda G_{\mathcal{F}}\sqrt{PM}L_g L_f a_k\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}_i^{(k-1)}\|.$$

By following a similar argument as in [27, Proposition 5.1] and by (A.24), (20), (33) and (34), one can show that

$$\begin{aligned}(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^{\star})^T\boldsymbol{g}(\boldsymbol{x}^{(k-1)}) \leq \frac{\bar{c}_k}{2a_k} + \frac{1}{2a_k}\left(\sum_{j=1}^{N}\|\boldsymbol{\lambda}_j^{(k-1)} - \boldsymbol{\lambda}\|^2 - \sum_{i=1}^{N}\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\|^2\right)\\ + 2N\sqrt{P}D_\lambda L_g\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\| + NC_g\|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\lambda}}^{\star}\|.\end{aligned} \tag{A.25}$$

By taking the weighted running average of (A.25), we obtain

$$\begin{aligned}(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^{\star})^T\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}) &\leq \frac{1}{A_k}\sum_{\ell=1}^{k}a_\ell(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^{\star})^T\boldsymbol{g}(\boldsymbol{x}^{(\ell-1)})\\ &\leq \frac{1}{2A_k}\sum_{\ell=1}^{k}\bar{c}_\ell + \frac{1}{2A_k}\left(\sum_{j=1}^{N}\|\boldsymbol{\lambda}_j^{(0)} - \boldsymbol{\lambda}\|^2 - \sum_{i=1}^{N}\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\|^2\right)\\ &\qquad + \frac{2N\sqrt{P}D_\lambda L_g}{A_k}\sum_{\ell=1}^{k}a_\ell\|\boldsymbol{x}^{(\ell-1)} - \hat{\boldsymbol{\alpha}}^{(\ell)}\| + \frac{NC_g}{A_k}\sum_{\ell=1}^{k}a_\ell\|\hat{\boldsymbol{\lambda}}^{(\ell-1)} - \hat{\boldsymbol{\lambda}}^{\star}\|\\ &\leq \frac{1}{2A_k}\sum_{\ell=1}^{k}\bar{c}_\ell + \frac{2ND_\lambda^2}{A_k} + \frac{2N\sqrt{P}D_\lambda L_g}{A_k}\sum_{\ell=1}^{k}a_\ell\|\boldsymbol{x}^{(\ell-1)} - \hat{\boldsymbol{\alpha}}^{(\ell)}\| + \frac{NC_g}{A_k}\sum_{\ell=1}^{k}a_\ell\|\hat{\boldsymbol{\lambda}}^{(\ell-1)} - \hat{\boldsymbol{\lambda}}^{\star}\| \qquad (A.26)\\ &\triangleq \xi^{(k-1)},\end{aligned}$$

where the first inequality is owing to the fact that $\boldsymbol{g}(\boldsymbol{x})$ is convex, and the last inequality is obtained by dropping $-\sum_{i=1}^{N}\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\|^2$ followed by applying (20). We claim that

$$\lim_{k\to\infty}\xi^{(k-1)} = 0. \tag{A.27}$$

To see this, note that the first and second terms in $\xi^{(k-1)}$ converge to zero as $k \to \infty$ since $\lim_{k\to\infty} A_k = \infty$ and $\sum_{\ell=1}^{\infty} \bar{c}_\ell < \infty$. The term $\frac{1}{A_k}\sum_{\ell=1}^{k} a_\ell \|\hat{\boldsymbol{\lambda}}^{(\ell-1)} - \hat{\boldsymbol{\lambda}}^\star\|$ also converges to zero since, by Lemma 6, $\lim_{k\to\infty} \|\hat{\boldsymbol{\lambda}}^{(k)} - \hat{\boldsymbol{\lambda}}^\star\| = 0$ and so does its weighted running average by [47, Lemma 3]. Similarly, the term $\frac{1}{A_k}\sum_{\ell=1}^{k} a_\ell \|\boldsymbol{x}^{(\ell-1)} - \hat{\boldsymbol{\alpha}}^{(\ell)}\|$ also converges to zero since $\lim_{k\to\infty} \|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\| = 0$ due to (A.15).

Now let $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^\star + \delta\, \frac{\left(\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)})\right)^+}{\|\left(\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)})\right)^+\|}$ which lies in $\mathcal{D}$, since $\|\boldsymbol{\lambda}\| \leq \|\hat{\boldsymbol{\lambda}}^\star\| + \delta \leq D_\lambda$ by (21). Substituting $\boldsymbol{\lambda}$ into (A.26) gives rise to

$$\delta \| \left(\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)})\right)^+ \| \leq \xi^{(k-1)}. \tag{A.28}$$

As a result, the first term in (A.23) is obtained by taking $k \to \infty$ in (A.28) and by (A.27).

To show that the second limit in (A.23) holds true, we first let $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^\star + \delta\, \frac{\hat{\boldsymbol{\lambda}}^{(k-1)}}{\|\hat{\boldsymbol{\lambda}}^{(k-1)}\|} \in \mathcal{D}$. By substituting it into (A.26) and by (20), we obtain $(\hat{\boldsymbol{\lambda}}^{(k-1)})^T \boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}) \leq \left(\frac{D_\lambda}{\delta}\right) \xi^{(k-1)}$ which, by taking $k \to \infty$, leads to

$$\limsup_{k\to\infty}\ (\hat{\boldsymbol{\lambda}}^{(k-1)})^T \boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}) \leq 0. \tag{A.29}$$

On the other hand, by letting $\boldsymbol{\lambda} = \boldsymbol{0} \in \mathcal{D}$, from (A.26) we have $-(\hat{\boldsymbol{\lambda}}^{(k-1)})^T \boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}) \leq \xi^{(k-1)} + (\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)})^T \boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}) \leq \xi^{(k-1)} + NC_g \|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\lambda}}^\star\|$. Since $\lim_{k\to\infty} \xi^{(k-1)} = 0$ and $\lim_{k\to\infty} \|\hat{\boldsymbol{\lambda}}^{(k)} - \hat{\boldsymbol{\lambda}}^\star\| = 0$ by Lemma 6, it follows that $\liminf_{k\to\infty}\ (\hat{\boldsymbol{\lambda}}^{(k-1)})^T \boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}) \geq 0$, which along with (A.29) yields the second term in (A.23). ■

By Lemma 6, Lemma 7 and Proposition 1, Theorem 2 is thus proved.

Appendix B

Proof of Theorem 3

Theorem 3 essentially can be obtained in the same line as the proof of Theorem 2, except for Lemma 5. What we need to show here is that the centralized proximal perturbation point $\hat{\boldsymbol{\alpha}}^{(k)}$ in (A.3) and $\hat{\boldsymbol{\beta}}^{(k)}$ in (A.2b) and the primal-dual iterates $(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)})$ satisfy a result similar to Lemma 5. The lemma below is proved in Appendix E:

Lemma 8 *Let Assumptions 2 and 3 hold. For the centralized perturbation points* $\hat{\boldsymbol{\alpha}}^{(k)}$ *in* (A.3) *and* $\hat{\boldsymbol{\beta}}^{(k)}$ *in* (A.2b)*, it holds true that*

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) &- \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \\ &\geq \left(\frac{1}{2\rho_1} - \frac{G_{\mathcal{F}}}{2}\right) \|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\|^2 + \frac{1}{\rho_2} \|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\|^2.\end{aligned} \tag{A.30}$$

Moreover, let $\rho_1 \leq 1/G_{\bar{\mathcal{F}}}$, and suppose that $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \to 0$ and $(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)})$ converges to some limit point $(\hat{\boldsymbol{x}}^{\star}, \hat{\boldsymbol{\lambda}}^{\star}) \in \mathcal{X} \times \mathcal{D}$ as $k \to \infty$. Then $(\hat{\boldsymbol{x}}^{\star}, \hat{\boldsymbol{\lambda}}^{\star})$ is a saddle point of (19).

Analogous to Lemma 6, as long as $\rho_1 \leq \frac{1}{G_{\bar{\mathcal{F}}}}$, the primal-dual iterates $(\boldsymbol{x}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ converge to a saddle point of (19).

Appendix C
Proof of Lemma 3

We first show (A.5). By definitions in (A.2b) and (23b), and by the non-expansiveness of projection, we readily obtain

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_i^{(k)}\| &\leq \left\| \mathcal{P}_{\mathcal{D}}\left(\tilde{\boldsymbol{\lambda}}_i^{(k)} + \rho_2 \ N\tilde{\boldsymbol{z}}_i^{(k)}\right) - \mathcal{P}_{\mathcal{D}}\left(\hat{\boldsymbol{\lambda}}^{(k-1)} + \rho_2 \ N\hat{\boldsymbol{z}}^{(k-1)}\right)\right\| \\ &\leq \|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + \rho_2 N \|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\|.\end{aligned}$$

Equation (A.4) for the $\boldsymbol{\alpha}_i^{(k)}$ in (23a) and $\hat{\boldsymbol{\alpha}}_i^{(k)}$ in (A.2a) can be shown in a similar line:

$$\begin{aligned}\|\boldsymbol{\alpha}_i^{(k)} - \hat{\boldsymbol{\alpha}}_i^{(k)})\| = &\left\| \mathcal{P}_{\mathcal{X}_i}\left(\boldsymbol{x}_i^{(k-1)} - \rho_1\left[\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\tilde{\boldsymbol{\lambda}}_i^{(k)}\right]\right)\right. \\ &\left. - \mathcal{P}_{\mathcal{X}_i}\left(\boldsymbol{x}_i^{(k-1)} - \rho_1\left[\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\hat{\boldsymbol{\lambda}}^{(k-1)}\right]\right)\right\| \\ \leq\ & \rho_1 \|\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\| \|\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) - \nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)})\| + \rho_1 \|\nabla \boldsymbol{g}_i(\boldsymbol{x}_i^{(k-1)})\| \|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| \\ \leq\ & \rho_1 L_g \sqrt{P} \|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + \rho_1 G_{\mathcal{F}} L_f \sqrt{M} N \|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\|, \end{aligned} \tag{A.31}$$

where, in the second inequality, we have used the boundedness of gradients (cf. (30), (32)) and the Lipchitz continuity of $\nabla\mathcal{F}$ (Assumption 3).

To show that (A.4) holds for $\boldsymbol{\alpha}_i^{(k)}$ in (24) and $\hat{\boldsymbol{\alpha}}_i^{(k)}$ in (A.3), we use the following lemma:

Lemma 9 [53, Lemma 4.1] *If $\boldsymbol{y}^{\star} = \arg\min_{\boldsymbol{y}\in\mathcal{Y}} J_1(\boldsymbol{y}) + J_2(\boldsymbol{y})$, where $J_1 : \mathbb{R}^n \to \mathbb{R}$ and $J_2 : \mathbb{R}^n \to \mathbb{R}$ are convex functions and $\mathcal{Y}$ is a closed convex set. Moreover, J_2 is continuously differentiable. Then $\boldsymbol{y}^{\star} = \arg\min_{\boldsymbol{y}\in\mathcal{Y}}\{J_1(\boldsymbol{y}) + \nabla J_2^T(\boldsymbol{y}^{\star})\boldsymbol{y}\}$.*

By applying the above lemma to (24) using $J_1(\boldsymbol{\alpha}_1) = \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\tilde{\boldsymbol{\lambda}}_i^{(k)}$ and

$$J_2(\boldsymbol{\alpha}_i) = \frac{1}{2\rho_1}\|\boldsymbol{\alpha}_i - (\boldsymbol{x}_i^{(k-1)} - \rho_1 \nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}))\|^2,$$

we obtain

$$\boldsymbol{\alpha}_i^{(k)} = \arg\min_{\boldsymbol{\alpha}_i \in \mathcal{X}_i} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\tilde{\boldsymbol{\lambda}}_i^{(k)} + (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) + \frac{1}{\rho_1}(\boldsymbol{\alpha}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}))^T\boldsymbol{\alpha}_i. \quad \text{(A.32)}$$

Similarly, applying Lemma 9 to (A.3), we obtain

$$\hat{\boldsymbol{\alpha}}_i^{(k)} = \arg\min_{\boldsymbol{\alpha}_i \in \mathcal{X}_i} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\hat{\boldsymbol{\lambda}}^{(k-1)} + (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)}) + \frac{1}{\rho_1}(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}))^T\boldsymbol{\alpha}_i. \quad \text{(A.33)}$$

From (A.32) it follows that

$$\begin{aligned} &\boldsymbol{g}_i^T(\boldsymbol{\alpha}_i^{(k)})\tilde{\boldsymbol{\lambda}}_i^{(k)} + (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) + \frac{1}{\rho_1}(\boldsymbol{\alpha}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}))^T\boldsymbol{\alpha}_i^{(k)} \\ &\quad \leq \boldsymbol{g}_i^T(\hat{\boldsymbol{\alpha}}_i^{(k)})\tilde{\boldsymbol{\lambda}}_i^{(k)} + (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) + \frac{1}{\rho_1}(\boldsymbol{\alpha}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}))^T\hat{\boldsymbol{\alpha}}_i^{(k)}, \end{aligned}$$

which is equivalent to

$$\begin{aligned} 0 \leq\ & (\boldsymbol{g}_i^T(\hat{\boldsymbol{\alpha}}_i^{(k)}) - \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i^{(k)}))\tilde{\boldsymbol{\lambda}}_i^{(k)} \\ &+ \nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)})(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{\alpha}_i^{(k)}) + \frac{1}{\rho_1}(\boldsymbol{\alpha}_i^{(k)} - \boldsymbol{x}_i^{(k-1)})(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{\alpha}_i^{(k)}). \end{aligned} \quad \text{(A.34)}$$

Similarly, equation (A.33) implies that

$$\begin{aligned} 0 \leq\ & (\boldsymbol{g}_i^T(\boldsymbol{\alpha}_i^{(k)}) - \boldsymbol{g}_i^T(\hat{\boldsymbol{\alpha}}_i^{(k)}))\hat{\boldsymbol{\lambda}}^{(k-1)} \\ &+ \nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)})(\boldsymbol{\alpha}_i^{(k)} - \hat{\boldsymbol{\alpha}}_i^{(k)}) + \frac{1}{\rho_1}(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)})(\boldsymbol{\alpha}_i^{(k)} - \hat{\boldsymbol{\alpha}}_i^{(k)}). \end{aligned} \quad \text{(A.35)}$$

By combining (A.34) and (A.35), we obtain

$$\begin{aligned} \frac{1}{\rho_1}\|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{\alpha}_i^{(k)}\|^2 &\leq (\boldsymbol{g}_i^T(\hat{\boldsymbol{\alpha}}_i^{(k)}) - \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i^{(k)}))(\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}) \\ &\quad + \nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})(\nabla\mathcal{F}(N\tilde{\boldsymbol{y}}_i^{(k)}) - \nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)}))(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{\alpha}_i^{(k)}) \\ &\leq \left(\sqrt{P}L_g\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + G_{\mathcal{F}}L_f\sqrt{M}N\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\|\right)\|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{\alpha}_i^{(k)}\|, \end{aligned}$$

where we have used the boundedness of gradients (cf. (30), (32)), the Lipschitz continuity of $\nabla\mathcal{F}$ (Assumption 3) as well as the Lipschitz continuity of $\boldsymbol{g}_i$ (in (33)). The desired result in (A.4) follows from the preceding relation. ■

Appendix D
Proof of Lemma 5

We first prove that relation (A.13) holds for the perturbation points $\hat{\boldsymbol{\alpha}}_i^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$ in (A.2) assuming that Assumption 4 is satisfied. Note that (A.2a) is equivalent to

$$\hat{\boldsymbol{\alpha}}_i^{(k)} = \arg\min_{\boldsymbol{\alpha}_i \in \mathcal{X}_i} \|\boldsymbol{\alpha}_i - \boldsymbol{x}_i^{(k-1)} + \rho_1 \mathcal{L}_{\boldsymbol{x}_i}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)})\|^2,\ i = 1, \ldots, N,$$

where $\mathcal{L}_{\boldsymbol{x}_i}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) = \nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)}) \nabla \mathcal{F}(N\hat{\boldsymbol{y}}^{(k)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\hat{\boldsymbol{\lambda}}^{(k-1)}$. By the optimality condition, we have that, for all $\boldsymbol{x}_i \in \mathcal{X}_i$,

$$(\boldsymbol{x}_i - \hat{\boldsymbol{\alpha}}_i^{(k)})^T(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)} + \rho_1 \mathcal{L}_{\boldsymbol{x}_i}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)})) \geq 0.$$

By choosing $\boldsymbol{x}_i = \boldsymbol{x}_i^{(k-1)}$, one obtains

$$(\boldsymbol{x}_i^{(k-1)} - \hat{\boldsymbol{\alpha}}_i^{(k)})^T \mathcal{L}_{\boldsymbol{x}_i}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \geq \frac{1}{\rho_1}\|\boldsymbol{x}_i^{(k-1)} - \hat{\boldsymbol{\alpha}}_i^{(k)}\|^2,$$

which, by summing over $i = 1, \ldots, N$, gives rise to

$$(\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)})^T \mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \geq \frac{1}{\rho_1}\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\|^2.$$

Further write the above equation as follows

$$\begin{aligned}
&(\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)})^T \mathcal{L}_{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \\
&\geq \frac{1}{\rho_1}\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\|^2 - (\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)})^T(\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - \mathcal{L}_{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})) \\
&\geq \frac{1}{\rho_1}\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\|^2 - \|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\| \times \|\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - \mathcal{L}_{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})\|. \qquad \text{(A.36)}
\end{aligned}$$

By (11), Assumption 3, Assumption 4 and the boundedness of $\hat{\boldsymbol{\lambda}}^{(k-1)} \in \mathcal{D}$, we can bound the second term in (A.36) as

$$\begin{aligned}
&\|\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - \mathcal{L}_{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})\| \\
&\leq \|\nabla \bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}) - \nabla \bar{\mathcal{F}}(\hat{\boldsymbol{\alpha}}^{(k)})\| + \|\hat{\boldsymbol{\lambda}}^{(k-1)}\| \left\| \begin{bmatrix} \nabla \boldsymbol{g}_1^T(\boldsymbol{x}_1^{(k-1)}) - \nabla \boldsymbol{g}_1^T(\hat{\boldsymbol{\alpha}}_1^{(k)}) \\ \vdots \\ \nabla \boldsymbol{g}_N^T(\boldsymbol{x}_N^{(k-1)}) - \nabla \boldsymbol{g}_N^T(\hat{\boldsymbol{\alpha}}_N^{(k)}) \end{bmatrix} \right\|_F \\
&\leq (G_{\bar{\mathcal{F}}} + D_\lambda \sqrt{P} G_g)\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\|, \qquad \text{(A.37)}
\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenious norm. By combining (A.36) and (A.37), we obtain

$$(\boldsymbol{x}^{(k-1)}-\hat{\boldsymbol{\alpha}}^{(k)})^T\mathcal{L}_{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)}) \geq \left(\frac{1}{\rho_1}-(G_{\bar{\mathcal{F}}}+D_\lambda\sqrt{P}G_g)\right)\|\boldsymbol{x}^{(k-1)}-\hat{\boldsymbol{\alpha}}^{(k)}\|^2. \tag{A.38}$$

Since $\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\lambda}}^{(k-1)})-\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)}) \geq (\boldsymbol{x}^{(k-1)}-\hat{\boldsymbol{\alpha}}^{(k)})^T\mathcal{L}_{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)})$ by the convexity of $\mathcal{L}$ in $\boldsymbol{x}$, we further obtain

$$\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\lambda}}^{(k-1)})-\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)}) \geq \left(\frac{1}{\rho_1}-(G_{\bar{\mathcal{F}}}+D_\lambda\sqrt{P}G_g)\right)\|\boldsymbol{x}^{(k-1)}-\hat{\boldsymbol{\alpha}}^{(k)}\|^2. \tag{A.39}$$

On the other hand, by (A.2b), we know that $\hat{\boldsymbol{\beta}}^{(k)}=\arg\min_{\beta\in\mathcal{D}}\|\boldsymbol{\beta}-\hat{\boldsymbol{\lambda}}^{(k-1)}-\rho_2\sum_{i=1}^N\boldsymbol{g}_i(\boldsymbol{x}_i^{(k-1)})\|^2$. By the optimality condition and the linearity of $\mathcal{L}$ in $\boldsymbol{\lambda}$, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})-\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\lambda}}^{(k-1)}) &= -(\hat{\boldsymbol{\lambda}}^{(k-1)}-\hat{\boldsymbol{\beta}}^{(k)})^T\left(\sum_{i=1}^N\boldsymbol{g}_i(\boldsymbol{x}_i^{(k-1)})\right)\\ &\geq \frac{1}{\rho_2}\|\hat{\boldsymbol{\lambda}}^{(k-1)}-\hat{\boldsymbol{\beta}}^{(k)}\|^2.\end{aligned} \tag{A.40}$$

Combining (A.39) and (A.40) yields (A.13).

Suppose that $\mathcal{L}(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\beta}}^{(k)})-\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)})\to 0$ and $(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\lambda}}^{(k-1)})$ converges to some limit point $(\hat{\boldsymbol{x}}^\star,\hat{\boldsymbol{\lambda}}^\star)$ as $k\to\infty$. Since $\rho_1\leq 1/(G_{\bar{\mathcal{F}}}+D_\lambda\sqrt{P}G_g)$, we infer from (A.13) that $\|\boldsymbol{x}^{(k-1)}-\hat{\boldsymbol{\alpha}}^{(k)}\|\to 0$ and $\|\hat{\boldsymbol{\lambda}}^{(k-1)}-\hat{\boldsymbol{\beta}}^{(k)}\|\to 0$, as $k\to\infty$. It then follows from (A.2) and the fact that projection is a continuous mapping [53] that $(\hat{\boldsymbol{x}}^\star,\hat{\boldsymbol{\lambda}}^\star)\in\mathcal{X}\times\mathcal{D}$ satisfies

$$\begin{aligned}&\hat{\boldsymbol{x}}_i^\star=\mathcal{P}_{\mathcal{X}_i}\left(\hat{\boldsymbol{x}}_i^\star-\rho_1[\nabla\boldsymbol{f}_i^T(\hat{\boldsymbol{x}}_i^\star)\nabla\mathcal{F}\left(\sum_{i=1}^M\boldsymbol{f}_i(\hat{\boldsymbol{x}}_i^\star)\right)+\nabla\boldsymbol{g}_i^T(\hat{\boldsymbol{x}}_i^\star)\hat{\boldsymbol{\lambda}}^\star]\right),\ i=1,\ldots,N,\\ &\hat{\boldsymbol{\lambda}}^\star=\mathcal{P}_{\mathcal{D}}(\hat{\boldsymbol{\lambda}}^\star+\rho_2\ \sum_{i=1}^N\boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^\star))\end{aligned}$$

which, respectively, imply that $\hat{\boldsymbol{x}}^\star=\arg\min_{\boldsymbol{x}\in\mathcal{X}}\mathcal{L}(\boldsymbol{x},\hat{\boldsymbol{\lambda}}^\star)$ and $\hat{\boldsymbol{\lambda}}^\star=\arg\max_{\boldsymbol{\lambda}\in\mathcal{D}}\mathcal{L}(\hat{\boldsymbol{x}}^\star,\boldsymbol{\lambda})$ i.e., $(\hat{\boldsymbol{x}}^\star,\hat{\boldsymbol{\lambda}}^\star)$ is a saddle point of problem (19). ■

APPENDIX E
PROOF OF LEMMA 8

The definition of $\hat{\boldsymbol{\alpha}}^{(k)}$ in (A.3) implies that

$$\begin{aligned}&\boldsymbol{g}_i^T(\hat{\boldsymbol{\alpha}}_i^{(k)})\hat{\boldsymbol{\lambda}}^{(k-1)}+(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)})^T\nabla\boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)})\\ &\qquad\qquad+\frac{1}{2\rho_1}\|\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\|^2\ \leq\ \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\hat{\boldsymbol{\lambda}}^{(k-1)},\end{aligned}$$

which, by summing over $i = 1, \ldots, N$, yields

$$\boldsymbol{g}^T(\hat{\boldsymbol{\alpha}}^{(k)})\hat{\boldsymbol{\lambda}}^{(k-1)} + (\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)})^T \nabla \bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}) \\ + \frac{1}{2\rho_1}\|\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)}\|^2 \ \leq \ \boldsymbol{g}^T(\boldsymbol{x}^{(k)})\hat{\boldsymbol{\lambda}}^{(k-1)}, \tag{A.41}$$

where $\boldsymbol{g}(\hat{\boldsymbol{\alpha}}^{(k)}) = \sum_{i=1}^N \boldsymbol{g}_i^T(\hat{\boldsymbol{\alpha}}_i^{(k)})$. By substituting the decent lemma in [53, Lemma 2.1]

$$\bar{\mathcal{F}}(\hat{\boldsymbol{\alpha}}^{(k)}) \leq \bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}) + (\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)})^T \nabla \bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}) + \frac{G_{\bar{\mathcal{F}}}}{2}\|\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)}\|^2 \tag{A.42}$$

into (A.41), we then obtain

$$\left(\frac{1}{2\rho_1} - \frac{G_{\bar{\mathcal{F}}}}{2}\right)\|\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)}\|^2 \leq \mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})$$

which, after combining with (A.40), yields (A.30).

To show the second part of this lemma, let us recall (A.33) that $\hat{\boldsymbol{\alpha}}_i^{(k)}$ in (A.3) can be alternatively written as

$$\hat{\boldsymbol{\alpha}}_i^{(k)} = \arg \min_{\boldsymbol{\alpha}_i \in \mathcal{X}_i} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\hat{\boldsymbol{\lambda}}^{(k-1)} + (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla \mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)}) + \frac{1}{\rho_1}(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}))^T \boldsymbol{\alpha}_i,$$

which implies that, for all $\boldsymbol{x}_i \in \mathcal{X}_i$, we have

$$\boldsymbol{g}_i^T(\hat{\boldsymbol{\alpha}}_i^{(k)})\hat{\boldsymbol{\lambda}}^{(k-1)} + (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla \mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)}) + \frac{1}{\rho_1}(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}))^T(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}) \\ \leq \ \boldsymbol{g}_i^T(\boldsymbol{x}_i)\hat{\boldsymbol{\lambda}}^{(k-1)} + (\nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla \mathcal{F}(N\hat{\boldsymbol{y}}^{(k-1)}) + \frac{1}{\rho_1}(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}))^T(\boldsymbol{x}_i - \boldsymbol{x}_i^{(k-1)}).$$

By summing the above inequality over $i = 1, \ldots, N$, one obtains, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\boldsymbol{g}^T(\hat{\boldsymbol{\alpha}}^{(k)})\hat{\boldsymbol{\lambda}}^\star + \nabla \bar{\mathcal{F}}^T(\boldsymbol{x}^{(k-1)})(\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)}) + \frac{1}{\rho_1}\|\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)}\|^2 \\ \leq \ \boldsymbol{g}^T(\boldsymbol{x})\hat{\boldsymbol{\lambda}}^\star + \nabla \bar{\mathcal{F}}^T(\boldsymbol{x}^{(k-1)})(\boldsymbol{x} - \boldsymbol{x}^{(k-1)}) \\ + \frac{1}{\rho_1}\sum_{i=1}^N (\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)})(\boldsymbol{x}_i - \boldsymbol{x}_i^{(k-1)}) + (\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)})(\boldsymbol{g}(\hat{\boldsymbol{\alpha}}^{(k)}) - \boldsymbol{g}(\boldsymbol{x})) \\ \leq \boldsymbol{g}^T(\boldsymbol{x})\hat{\boldsymbol{\lambda}}^\star + \bar{\mathcal{F}}(\boldsymbol{x}) - \bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)}) + \frac{2D_x}{\rho_1}\sum_{i=1}^N \|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\| + 2C_g\|\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)}\|,$$

where we have utilized the convexity of $\bar{\mathcal{F}}$ (Assumption 3), boundedness of $\mathcal{X}_i$ and the constraint functions (cf. Assumption 2 and (34)) in obtaining the last inequality. By applying (A.42) to the above inequality and by the premise of $1/\rho_1 \geq G_{\bar{\mathcal{F}}} > G_{\bar{\mathcal{F}}}/2$, we further obtain, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \leq \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}^\star) + \frac{2D_x}{\rho_1} \sum_{i=1}^{N} \|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\| \\ + 2C_g \|\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + |\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^\star)|, \tag{A.43}$$

in which one can bound the last term, using (38), (31), (33) and (20), by

$$|\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^\star)| \leq (L_{\bar{\mathcal{F}}} + N D_\lambda \sqrt{P} L_g) \|\hat{\boldsymbol{x}}^\star - \hat{\boldsymbol{\alpha}}^{(k)}\|. \tag{A.44}$$

Suppose that $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \to 0$ and $(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)})$ converges to some limit point $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)$ as $k \to \infty$. Then, by (A.30) and since $1/\rho_1 \geq G_{\bar{\mathcal{F}}}$, we have $\|(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - (\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})\| \to 0$, as $k \to \infty$. Therefore,

$$\lim_{k \to \infty} \left(\frac{2D_x}{\rho_1} \sum_{i=1}^{N} \|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\| + 2C_g \|\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)}\| + |\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^\star) - \mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)| \right) = 0.$$

Thus, it follows from (A.43), (A.44) and the above equation that $\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \leq \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}^\star)$ for all $\boldsymbol{x} \in \mathcal{X}$. The rest of the proof is similar to that of Lemma 5. ■

REFERENCES

[1] V. Lesser, C. Ortiz, and M. Tambe, *Distributed Sensor Networks: A Multiagent Perspective*. Kluwer Academic Publishers, 2003.

[2] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proc. ACM IPSN*, Berkeley, CA, USA, April 26-27, 2004, pp. 20–27.

[3] M. Chiang, P. Hande, T. Lan, and W. C. Tan, "Power control in wireless cellular networks," *Foundations and Trends in Networking*, vol. 2, no. 4, pp. 381–533, 2008.

[4] S. Chao, T.-H. Chang, K.-Y. Wang, Z. Qiu, and C.-Y. Chi, "Distributed robust multicell coordianted beamforming with imperfect csi: An ADMM approach," *IEEE Trans. Signal Processing*, vol. 60, no. 6, pp. 2988–3003, 2012.

[5] D. Belomestny, A. Kolodko, and J. Schoenmakers, "Regression methods for stochastic control problems and their convergence analysis," *SIAM J. on Control and Optimization*, vol. 48, no. 5, pp. 3562–3588, 2010.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer-Verlag, 2001.

[7] M. Elad, *Sparse and Redundant Rerpesentations*. New York, NY, USA: Springer Science + Business Media, 2010.

[8] R. Cavalcante, I. Yamada, and B. Mulgrew, "An adaptive projected subgradient approach to learning in diffusion networks," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2762–2774, Aug. 2009.

[9] B. Johansson, T. Keviczky, M. Johansson, and K. Johansson, "Subgradient methods and consensus algorithms for solving convex optimization problems," in *Proc. IEEE CDC*, Cancun, Mexica, Dec. 9-11, 2008, pp. 4185–4190.

[10] A. Nedić, A. Ozdaglar, , and A. Parrilo, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[11] ——, “Constrained consensus and optimization in multi-agent networks,” *IEEE Trans. Automatic Control*, vol. 55, no. 4, pp. 922–938, April 2010.

[12] I. Lobel and A. Ozdaglar, “Distributed subgradient methods for convex optimization over random networks,” *IEEE Trans. Automatic Control*, vol. 56, no. 6, pp. 1291–1306, June 2011.

[13] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Distributed stochastic subgradeint projection algorithms for convex optimization,” *J. Optim. Theory Appl.*, vol. 147, pp. 516–545, 2010.

[14] J. Chen and A. H. Sayed, “Diffusion adaption strategies for distributed optimization and learning networks,” *IEEE. Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, month =Aug., year = 2012.

[15] M. Zhu and S. Martínez, “On distributed convex optimization under inequality and equality constraints,” *IEEE Trans. Automatic Control*, vol. 57, no. 1, pp. 151–164, Jan. 2012.

[16] D. Yuan, S. Xu, and H. Zhao, “Distributed primal-dual subgradient method for multiagent optimization via consensus algorithms,” *IEEE Trans. Systems, Man, and Cybernetics- Part B*, vol. 41, no. 6, pp. 1715–1724, Dec. 2011.

[17] S. S. Ram, A. Nedić, and V. V. Veeravalli, “A new class of distributed optimization algorithm: Application of regression of distributed data,” *Optimization Methods and Software*, vol. 27, no. 1, pp. 71–88, 2012.

[18] T.-H. Chang, M. Alizadeh, and A. Scaglione, “Coordinated home energy management for real-time power balancing,” in *Proc. IEEE PES General Meeting*, San Diego, CA, July 22-26, 2012, pp. 1–8.

[19] N. Li, L. Chen, and S. H. Low, “Optimal demand response based on utility maximization in power networks,” in *IEEE PES General Meeting*, Detroit, MI, USA, July 24-29, 2011, pp. 1–8.

[20] J. C. V. Quintero, “Decentralized control techniques applied to electric power distributed generation in microgrids,” MS thesis, Departament d’Enginyeria de Sistemes, Automática i Informática Industrial, Universitat Politécnica de Catalunya, 2009.

[21] M. Kallio and A. Ruszczyński, “Perturbation methods for saddle point computation,” 1994, report No. WP-94- 38, International Institute for Applied Systems Analysis.

[22] M. Kallio and C. H. Rosa, “Large-scale convex optimization via saddle-point computation,” *Oper. Res.*, pp. 93–101, 1999.

[23] A. Olshevsky and J. N. Tsitsiklis, “Convergence rates in distributed consensus averaging,” in *Proc. IEEE CDC*, San Diego, CA, USA, Dec. 13-15, 2006, pp. 3387–3392.

[24] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” *Systems and Control Letters*, vol. 53, pp. 65–78, 2004.

[25] B. Yang and M. Johansson, “Distributed optimization and games: A tutorial overview,” Chapter 4 of *Networked Control Systems*, A. Bemporad, M. Heemels and M. Johansson (eds.), LNCIS 406, Springer-Verlag, 2010.

[26] H. Uzawa, “Iterative methods in concave programming,” 1958, in Arrow, K., Hurwicz, L., Uzawa, H. (eds.) Studies in Linear and Nonlinear Programming, pp. 154-165. Stanford University Press, Stanford.

[27] A. Nedić and A. Ozdaglar, “Subgradeint methods for saddle-point problems,” *J. OPtim. Theory Appl.*, vol. 142, pp. 205–228, 2009.

[28] K. Srivastava, A. Nedić, and D. Stipanović, “Distributed Bregman-distance algorithms for min-max optimization,” in book *Agent-Based Optimization*, I. Czarnowski, P. Jedrzejowicz and J. Kacprzyk (Eds.), Springer Studies in Computational Intelligence (SCI), 2012.

[29] M. Alizadeh, X. Li, Z. Wang, A. Scaglione, and R. Melton, “Demand side management in the smart grid: Information processing for the power switch,” *IEEE Signal Process. Mag.*, vol. 59, no. 5, pp. 55–67, Sept. 2012.

[30] X. Guan, Z. Xu, and Q.-S. Jia, “Energy-efficient buildings facilitated by microgrid,” *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 243–252, Dec. 2010.

[31] N. Cai and J. Mitra, “A decentralized control architecture for a microgrid with power electronic interfaces,” in *Proc. North American Power Symposium (NAPS)*, Sept. 26-28 2010, pp. 1–8.

[32] D. Hershberger and H. Kargupta, “Distributed multivariate regression using wavelet based collective data mining,” *J. Parallel Distrib. Comput.*, vol. 61, no. 3, pp. 372–400, March 2001.

[33] H. Kargupta, B.-H. Park, D. Hershberger, and E. Johnson, “Collective data mining: A new perspective toward distributed data mining,” in *Advances in Distributed Data Mining*, H. Kargupta and P. Chan (eds.), AAAI/MIT Press, 1999.

[34] A. Sanil, A. F. Karr, X. Lin, and J. P. Reiter, “Privacy preserving regression modelling via distributed computation,” *Proc. Int. Conf. Knowledge Discovery and Data Mining*, Seattle, WA, 2004, pp. 677-682.

[35] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Puschel, “Distributed basis pursuit,” *IEEE. Trans. Signal Process.*, vol. 60, no. 4, pp. 1942–1956, April 2012.

[36] P. Georgiev, F. Theis, A. Cichocki, and H. Bakardjian, “Sparse component analysis: A new tool for data mining,” *Data Mining in Biomedicine*, Springer Optimization and Its Applications, P. Pardalos. and V. Boginski and A. Vazacopoulos (eds.), vol. 7, pp. 91-116, 2007.

[37] L. F. M. Fardad and M. R. Jovanovic, “Sparsity-promoting optimal control for a class of distributed systems,” in *Proc. American Control Conference (ACC)*, San Francisco, CA, USA, June 29 - July 1, 2011, pp. 2050–2055.

[38] M. Nagahara, D. E. Quevedo, J. Ostergaard, T. Matsuda, and K. Hayashi, “Sparse command generator for remote control,” in *Proc. IEEE Int. Conf. on Control and Automation (ICCA)*, Santiago, Chile, Dec. 19-21, 2011, pp. 1055–1059.

[39] D. P. Bertsekas, *Network Optimization : Contribuous and Discrete Models.* Athena Scientific, 1998.

[40] R. Madan and S. Lall, “Distributed algorithms for maximum lifetime routing in wireless sensor networks,” *IEEE. Trans. Wireless Commun.*, vol. 5, no. 8, pp. 2185–2193, month =Aug., year = 2006.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[42] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*. Cambridge, Massachusetts: Athena Scientific, 2003.

[43] I. Y. Zabotin, “A subgradient method for finding a saddle point of a convex-concave function,” *Issled. Prikl. Mat.*, vol. 15, pp. 6–12, 1988.

[44] Y. Nesterov, “Smooth minimization of nonsmooth functions,” *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[45] G. Mateos, J. A. Bazerque, and G. B. Giannakis, “Distributed sparse linear regression,” *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5262–5276, Dec. 2010.

[46] A. Nedić and A. Ozdaglar, “Approximate primal solutions and rate analysis for dual subgradient methods,” *SIAM J. OPtim.*, vol. 19, no. 4, pp. 1757–1780, 2009.

[47] T. Larsson, . Patriksson, and A.-B. Strömberg, “Ergodic, primal convergence in dual subgradient schemes for convex programming,” *Math. Program.*, vol. 86, pp. 238–312, 1999.

[48] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 1.21,” http://cvxr.com/cvx, Apr. 2011.

[49] T.-H. Chang, M. Alizadeh, and A. Scaglione, “Real-time power balancing via decentralized coordinated home energy scheduling,” to appear in *IEEE Trans. Smart Grid*, 2013.

[50] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Asynchronous gossip algorithms for stochastic optimization,” in *Proc. Int. Conf. on Game Thoery for Networks*, Istanbul, Turkey, May 13-15, 2009, pp. 1–6.

[51] A. Nedić, “Asynchronous broadcast-based convex optimization over a network,” *IEEE Trans. Automatic Control*, vol. 56, no. 6, pp. 1337–1351, June 2011.

[52] B. T. Polyak, *Introduction to Optimization*. New Yoir: Optimization Software Inc., 1987.

[53] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: Numerical methods.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1989.